On the Assumption of Bivariate Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence

McGovern, M. E., Bärnighausen, T., Marra, G., & Radice, R. (2015). On the Assumption of Bivariate Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence. *Epidemiology*, *26*(2), 229-237. https://doi.org/10.1097/EDE.0000000000000218

Published in:
Epidemiology

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2015 Epidemiology
This is the final accepted manuscript and not the final published version. The final version can be found at http://journals.lww.com/epidem/pages/articleviewer.aspx?year=2015&issue=03000&article=00016&type=abstract

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Download date:06. May. 2023

On the Assumption of Bivariate Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence*

Mark E. McGovern[†1], Till Bärnighausen[2,3], Giampiero Marra[4], and Rosalba Radice[5]

[1]Harvard University
[2]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, USA
[3]Wellcome Trust Africa Centre for Health and Population Studies, University of KwaZulu-Natal, South Africa
[4]Department of Statistical Science, University College London
[5]Department of Economics, Mathematics and Statistics, Birkbeck, University of London

March 2015

Abstract

Heckman–type selection models have been used to control HIV prevalence estimates for selection bias when participation in HIV testing and HIV status are associated after controlling for observed variables. These models typically rely on the strong assumption that the error terms in the participation and the outcome equations that comprise the model are distributed as bivariate normal. We introduce a novel approach for relaxing the bivariate normality assumption in selection models using copula functions. We apply this method to estimating HIV prevalence and new confidence intervals (CI) in the 2007 Zambia Demographic and Health Survey (DHS) by using interviewer identity as the selection variable that predicts participation (consent to test) but not the outcome (HIV status). We show in a simulation study that selection models can generate biased results when the bivariate normality assumption is violated. In the 2007 Zambia DHS, HIV prevalence estimates are similar irrespective of the structure of the association assumed between participation and outcome. For men, we estimate a population HIV prevalence of 21% (95% CI = 16% – 25%) compared with 12% (11% – 13%) among those who consented to be tested; for women, the corresponding figures are 19% (13% – 24%) and 16% (15% – 17%). Copula approaches to Heckman–type selection models are a useful addition to the methodological toolkit of HIV epidemiology and of epidemiology in general. We develop the use of this approach to systematically evaluate the robustness of HIV prevalence estimates based on selection models, both empirically and in a simulation study.

JEL Classification: *C34, J11*

Keywords: *Selection Bias, Bivariate Normality, Copula, HIV*

*Published as: McGovern, M.E., Bärnighausen, T., Marra, G., Radice, R., 2015. On the Assumption of Bivariate Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence. Epidemiology 26, 229 – 327. Available at http://dx.doi.org/10.1097/EDE.0000000000000218. We are grateful to Slawa Rokicki for invaluable suggestions and advice. We also thank the editor, two anonymous referees, seminar participants at Harvard University, University College London, and the UNAIDS Reference Group on Estimates, Modelling, and Projections for comments. Mark McGovern received financial support from the Program on the Global Demography of Aging which receives funding from the National Institute on Aging (grant no. 1 P30 AG 024409-09). Till Bärnighausen received financial support through grant 1 R01 HD 058482 01 from the National Institute of Child Health and Human Development, National Institutes of Health.

[†]Corresponding author. Current Address: Queen's University Belfast, Riddel Hall, 185 Stranmillis Road, Belfast, BT9 5EE, Northern Ireland. Email: markemcgovern@gmail.com.

1 Introduction

To address almost every aspect of the HIV epidemic, from assessing the risk factors associated with infection to planning future resource allocation to antiretroviral treatment scale-up, accurate information on HIV prevalence is required (Gersovitz, 2011). Research on HIV often relies on nationally representative surveys (Boerma et al., 2003), but participation rates in these surveys can be low. Table 1 shows that participation rates in the HIV surveys that are nested within one of the major sources of nationally representative data in low- and middle-income countries, the Demographic and Health Surveys (DHS), range from a high of 97% for women in Rwanda in 2005 to a low of 63% for men in Malawi in 2004 and Zimbabwe in 2005 (Hogan et al., 2012). There are many potential reasons for low participation rates in HIV surveys (including concerns about the confidentiality of results, lack of incentives to participate, and survey fatigue) (Gersovitz, 2011; Sterck, 2013), and non-participation can arise at different stages of HIV survey administration (Marston et al., 2008). In this article, we focus on refusal to be tested for HIV, which is typically the most important cause of missing data in HIV surveys (Mishra et al., 2008). In longitudinal studies, it has been shown that respondents who are HIV positive are less likely to consent to be tested for HIV than HIV-negative individuals (Bärnighausen et al., 2012; Floyd et al., 2013; Obare, 2010; Reniers and Eaton, 2009). In Malawi, 46% of women and 39% of men who declined to be tested did so because of prior HIV testing, knowledge of HIV status, or fear of positive results (Kranzer et al., 2008). Such reasons for declining to participate in an HIV survey have implications for the estimation of HIV prevalence. Neither complete case analysis (limiting the analysis only to people who consent to be tested for HIV) nor standard approaches to account for missing values generate unbiased estimates in the presence of selection on unobserved variables (Conniffe and O'Neill, 2011; Donders et al., 2006).

A potentially important situation leading to selection into survey participation based on unobserved variables occurs if HIV status itself predicts consent to be tested for HIV. This situation is likely if people know (or correctly predict) that they are HIV positive and fear that others will learn about their positive HIV status if they participate in a survey. In this case, standard approaches to correct HIV prevalence estimates for missing values (such as single imputation, multiple imputation, inverse probability weighting, or propensity score reweighting) will lead to biased results because these approaches can only account for selection on observed factors, but HIV status is unobserved among those who refused to be tested. Another consequence of high refusal rates is that the uncertainty associated with estimating HIV prevalence can increase substantially, leading to wide confidence intervals (CI)(Hogan et al., 2012). More generally, missing

Table 1: Participation Rates for HIV Testing in Demographic and Health Surveys

	Participation Rates	
Demographic and Health Survey	Men (%)	Women (%)
Cote d'Ivoire 2005	76	79
Malawi 2004	63	70
Tanzania 2003	77	84
Tanzania 2007	80	90
Zimbabwe 2005	63	76
Lesotho 2004	68	81
Liberia 2007	81	88
Sierra Leone 2008	87	90
Zambia 2007	72	77
Cameroon 2004	90	92
Ethiopia 2005	76	83
Mali 2006	85	93
Niger 2006	84	91
Senegal 2005	75	84
Swaziland 2006	78	87
Rwanda 2005	96	97
Burkina Faso 2003	86	92
Congo 2007	86	90
Ghana 2003	80	89
Guinea 2005	88	92
Kenya 2003	70	76
Kenya 2008	79	86
Mali 2001	76	85
Zambia 2001	73	79

Source: Hogan et al. (2012). Data are publicly available from www.dhsprogram.com.

data are a common problem in epidemiologic studies, and the mechanisms through which this occurs can have an important impact on resulting estimates. Heckman–type selection models can provide asymptotically unbiased estimates of the parameters of interest, even when missing data are systematically related to unobserved characteristics of the individual (Heckman, 1979; Vella, 1998). These models will thus be useful whenever researchers cannot be certain that the assumption that is required for the standard approaches to generate unbiased results holdsthat is, that data are missing at random after selection on observed variables has been taken into account. However, in practice, the use of Heckman–type selection models is limited by one requirement and one main statistical assumption. Heckman–type selection models require the existence of a selection variable that predicts participation in a survey but not the outcome of interest, other than through the effect on participation. Elements of survey design and implementation are often documented in datasets in epidemiology (O'Muircheartaigh and Campanelli, 1998). Characteristics of these elements are often likely to determine survey participation and are thus potential candidates for selection variables if they are also plausibly uncorrelated with the characteristics of the individuals who are potential participants in a survey (Bärnighausen et al., 2011b). In HIV surveys, interviewer identity generally predicts consent to be tested, but it is unlikely that it also predicts HIV status. Previous research that has used interviewer identity as a selection variable in Heckman–type selection models has found evidence for selection on unobserved variables in several HIV surveys (Hogan et al., 2012; Bärnighausen et al., 2011a; Janssens et al., 2014; McGovern et al., 2015; Reniers et al., 2009). The key statistical assumption that the standard Heckman–type selection models need to meet is that the relation between consenting to be tested for HIV and HIV status follows a bivariate normal distribution after other covariates have been taken into account, that is, that the error terms of the 2 equations in Heckman–type selection models are distributed as bivariate normal. Although this assumption is convenient and tractable, it is a potentially serious limitation (Arpino et al., 2014; Geneletti et al., 2011; Puhani, 2000). If this assumption is met, then the estimates obtained using the conventional bivariate probit Heckmantype selection model are consistent and asymptotically efficient. However, if the true distribution of the error terms is not bivariate normal, then the estimates are likely to be both inconsistent and inefficient (De Luca, 2008). Simulation studies have indicated that HIV prevalence estimates from selection models may indeed be sensitive to violations of this assumption (Clark and Houle, 2012). The robustness of results obtained from surveys involving missing data is particularly important (Geneletti et al., 2011). The implementation of selection models can be viewed as a sensitivity analysis to adjust for potential bias using alternative sets of assumptions about the underlying mechanisms causing data to be missing. If it can be demonstrated that the results obtained in selection models are invariant to

a variety of alternative assumptions regarding the mechanisms leading to missing data, our belief that the conclusions are not just a function of the model imposed by the researcher will be substantially strengthened. The lack of methods for evaluating the robustness of Heckman selection models is likely an impediment to wider use of this approach.

The aim of this article is to develop and illustrate a means of determining the sensitivity of results from selection models with binary outcomes to alternative ways of characterizing the functional form of the association between the participation equation (in this case, consent to be tested for HIV) and the outcome equation (in this case, HIV status). Copulae have been previously applied to recursive models involving a treatment that is affected by unobserved variables (such as health as function of medical care utilization) (Dancer et al., 2008; Murteira and Lourenço, 2011; Prieger, 2002; Winkelmann, 2012), and in censored models with continuous outcomes (Smith, 2003). The main contribution of this article is the application of copulae to binary outcomes with missing data. In addition, we use a variety of copulae (including the rotated Clayton, Joe, and Gumbel), allowing for more flexibility in modelling dependence. This flexibility is a key characteristic of our approach because it allows us to capture a much wider set of possible dependence structures than those used in the previous literature (Radice et al., 2015). With this method, and the number of alternative parametric specifications, we are therefore able to be more confident in assessing the robustness of results based on the standard Heckman–type selection models. For example, whereas previous implementations of the copula approach have generally focused on distributions that are similar to the bivariate normal (such as the Frank) (Winkelmann, 2012), we are able to consider asymmetric dependence. In addition to potential bias arising from misspecification of the error distribution, by potentially providing a more accurate representation of the underlying data structure, the copula approach may also provide more efficient estimates, allowing us to make better inferences. The proposed approach has not been previously implemented in the sample selection literature. In what follows, we introduce and demonstrate our methodology for relaxing the assumption of bivariate normality in Heckman–type selection models that allow for nonlinear association between participation and the outcome of interest. Although, in theory, semiparametric or nonparametric approaches would not require any distributional assumptions, their application to estimating the intercept in sample selection models with binary data and a high degree of missing data is limited due to their inefficiency and computational feasibility. Although the copula method does require parametric specification, our approach makes many distributional functional forms available, therefore making copulae a viable practical alternative to imposing bivariate normality. We illustrate the consequences of violating the normality assumption in a simulation study and show that copulae can provide an effective and practical

means of adjusting for this bias and inefficiency. Finally, we evaluate the robustness of estimates of HIV prevalence in Zambia. We provide the relevant code to make this approach easily accessible to researchers working with surveys containing missing data (eAppendix; http://links.lww.com/EDE/A858; and online at http://dx.doi.org/10.7910/DVN/27727).

2 Methods

2.1 Statistical Approach

We begin by modelling consent to be tested for HIV and HIV status simultaneously, an approach based on the adaptation of the original Heckman selection model estimator for binary outcomes (Dubin and Rivers, 1989; Heckman, 1979; Van de Ven and Van Praag, 1981). Consent to be tested is given by:

$$Consent_i^* = X_i^T \beta_c + Z_i^T \alpha + u_i, \ i = 1...n \tag{1}$$

$$Consent_i = 1 \text{ if Consent}_i^* > 0, \ \text{Consent}_i = 0 \text{ otherwise.} \tag{2}$$

The observed consent for person i, $Consent_i$, is a dummy variable indicating acceptance of being tested and is a function of a latent variable, $Consent_i^*$, which reflects the respondent's propensity to be tested. X_i is a $p \times 1$ vector representing observed individual level characteristics with associated parameter vector β, Z_i is a $k \times 1$ vector of dummy variables representing interviewer identity with associated parameter vector α, and u_i is a random error term. T denotes the matrix transpose function. Although, in theory, identification can be achieved using the same set of regressors in both the participation equation and the outcome equation, in practice, empirical identification in selection models requires at least one variable, the selection variable, to be present in the participation equation but not the outcome equation (Madden, 2008; Smith, 2003). In this case, interviewer identity predicts consent to be tested but does not to enter into the HIV equation directly.

The equation for the HIV status HIV_i of individual i is:

$$HIV_i^* = X_i^T \gamma + \epsilon_i, \tag{3}$$

$$HIV_i = 1 \text{ if } \text{HIV}_i^* > 0,\ \text{HIV}_i = 0 \text{ otherwise,} \tag{4}$$

$$\text{HIV}_i \text{ observed only if } \text{Consent}_i = 1,\ \text{missing otherwise,} \tag{5}$$

where γ is a parameter vector and ϵ_i is a random error term. The structural assumption used in previous studies to estimate HIV prevalence is that the error terms in both equations (u_i, ϵ_i) are independent and identically distributed as bivariate normal, with means equal to zero, constant variances equal to 1, and covariance (correlation coefficient) ρ. That is, the joint cumulative distribution function (cdf) of (u_i, ϵ_i) is given by $F(u_i, \epsilon_i) = \phi_2(u_i, \epsilon_i, \rho)$, where ϕ_2 is the standardized bivariate cdf. This model can be fitted using classic maximum likelihood. The standard selection model that relies on joint normality is equivalent to specifying the Gaussian copula in our framework; therefore, we use this model as the baseline for our comparisons. To allow for nonlinear association between the consent and HIV status equations, we model the dependency of the error terms in the 2 equations using copulae. Broadly speaking, these are functions that connect multivariate distributions to their 1D margins, such that if F is a 2D cdf with 1D margins $(F_1(y_1), F_2(y_2))$, then there exists a 2D copula C such that $F(y_1, y_2) = C(F_1(y_1), F_2(y_2); \theta)$, where y_1 and y_2 are 2 random variables, and θ is an association parameter measuring the dependence between the 2 marginals (Trivedi and Zimmer, 2007).

If HIV positive persons are refusing to be tested on the basis of knowledge of their HIV status (Bärnighausen et al., 2012; Floyd et al., 2013; Obare, 2010; Reniers and Eaton, 2009), we would expect a value of $\theta < 0$. When we estimate the model for Zambia using copulae (Gaussian, Frank, and Student-t) that do not impose a sign on the relation between consent and HIV status, the dependence is estimated to be negative in the data, and when we implement copulae that specify positive associations (Clayton 0 and 180, Joe 0 and 180, and Gumbel 0 and 180), we find that the models do not converge. Therefore, we focus on those copulae that allow for negative association. However, in other contexts, there could just as easily be a positive relation, when this method is equally applicable. The models we consider are therefore: Gaussian (C_g), equivalent to the standard bivariate normal probit model; Frank (C_f); 90 and 270 degrees rotated Clayton ($C_{C_{90}}$, $C_{C_{270}}$); 90 and 270 degrees rotated Joe ($C_{J_{90}}$, $C_{J_{270}}$); 90 and 270 degrees rotated Gumbel ($C_{G_{90}}$, $C_{G_{270}}$); and Student–t (C_t). These copulae are listed in Table 2 and illustrated in Figure 1 (we omit the Joe copulae from Figure 1 as they are similar in appearance to the corresponding Clayton and Gumbel versions). The rotated

Clayton, Joe, and Gumbel copulae allow for stronger negative dependence in the tails of the distribution. The 90- and 270-degrees rotated versions can be obtained using the following equations (Brechmann and Schepsmeier, 2012):

$$C_{90} = F_2(y_2) - C(1 - F_1(y_1), F_2(y_2); \theta),$$

$$C_{270} = F_1(y_1) - C(F_1(y_1), 1 - F_2(y_2); \theta).$$

These forms of dependence are particularly applicable in the context of HIV prevalence estimation, as we might expect respondents with a strong negative score on the latent test variable to be of particularly high risk of being HIV positive. In the sample selection context, the data identify the 3 possible events $(Consent_i = 1, HIV_i = 1), (Consent_i = 1, HIV_i = 0)$, and $(Consent_i = 0)$, with probabilities:

$$P(Consent_i = 1, HIV_i = 1) = p_{11i} = C(\Phi(X_i^T\beta + Z_i^T\alpha), \Phi(X_i^T\gamma); \theta),$$

$$P(Consent_i = 1, HIV_i = 0) = p_{01i} = \Phi(X_i^T\beta + Z_i^T\alpha) - p_{11i},$$

$$P(Consent_i = 0) = p_{0i} = 1 - \Phi(X_i^T\beta + Z_i^T\alpha).$$

The log-likelihood function is therefore:

$$\sum_{i=1}^{n} Consent_i \times HIV_i log(p_{11i}) + Consent_i \times (1 - HIV_i) log(p_{01i}) + (1 - Consent_i) log(p_{0i}) \tag{6}$$

where $\delta^T = (\beta^T, \alpha^T, \gamma^T, \theta)$.

Maximization is based on a trust region algorithm and not the usual Newton–Raphson algorithm, resulting in more stable computation and better convergence properties, which is valuable because another common criticism of these models is that they can often fail to converge.

We roughly assess the degree of association between the consent and HIV status equations using a nonparametric measure of rank (Kendall's tau, τ). τ can be interpreted in the same manner as ρ in the sense that it ranges between 1 and +1; therefore, if persons who refuse to be tested are more likely to be HIV positive, we would expect to see a value of $\tau < 0$. The approximate posterior cdf $\hat{F}(\tau)$ is obtained by simulating a set of random values, $\{\theta_r : r = 1, ..., R\}$, from the multivariate normal posterior of δ such that:

$$\hat{F}(\tau) = \frac{1}{R} \sum_{r=1}^{R} H(\tau - \tau(\theta_r))$$

where H is the Heaviside function (jumping from 0 to 1 at τ). CI are obtained from quantiles of this distribution. Intervals for $\tau(\theta)$ may also be obtained by bootstrapping. The HIV prevalence estimate is computed as a weighted average of individual predicted values with survey weights, w_i:

$$\hat{P}(HIV = 1) = \frac{(\sum_{i=1}^{n} w_i \hat{P}(HIV = 1|X_i))}{\sum_{i=1}^{n} w_i}$$

We use a Taylor–series expansion to derive the large–sample variance estimator for the point estimate of HIV prevalence, which simultaneously acknowledges uncertainty due to cluster effects and the presence of sampling weights (Lumley, 2004).

There are no disadvantages to not specifying the standard normality assumption as we are using a likelihood–based model; hence, asymptotic theory will still hold under the usual regularity conditions, and we can evaluate model fit using information criteria (eg, the Akaike Information Criterion [AIC]). However, it is important to understand the relative performance of the standard Heckman–type model in comparison with the copula approach. Therefore, we undertake a simulation study to determine the conditions under which the normality assumption performs well and to assess the extent of bias that arises from misspecification of the error terms' distribution.

2.2 Simulation Study

We follow the approach implemented in a study by Clark and Houle (2012) by generating a dataset based on a real HIV survey (the 2007 Zambia DHS). Therefore, our simulations closely match the observed consent rates and HIV prevalence in the data used in the empirical part of this article. We construct latent variables

Table 2: Definition of Copula Functions

Copula	$C(F_1(y1), F_2(y_2); \theta)$
Normal	$\phi_2(\phi^{-1}(F_1), \phi^{-1}(F_2); \theta)$
Frank	$-\theta^{-1} ln(1 + \frac{(e^{-\theta F_1}-1)(e^{-\theta F_2}-1)}{(e^{-\theta}-1)})$
Clayton	$(F_1^{-\theta} + F_2^{-\theta} - 1)^{\frac{-1}{\theta}}$
Student	$t_{2v}(t_v^{-1}(F_1), t_v^{-1}(F_2); \theta)$
Joe	$1 - ((1-F_1)^\theta + (1-F_2)^\theta - (1-F_1)^\theta(1-F_2)^\theta)^{\frac{1}{\theta}}$
Gumbel	$\exp(-((-log(F_1))^\theta + (-log(F_2))^\theta)^{\frac{1}{\theta}}$

Note: $t_{2v}(.,.;\theta)$ denotes the cumulative distribution function of a standard bivariate Student-t distribution with correlation coefficient θ and v degrees of freedom. $t_v^{(-1)}$ denotes the inverse univariate Student-t distribution function with v degrees of freedom.

for consent and HIV status and allow for interviewer identity to influence the probability of consent. Then we draw error terms for the latent variable equations to induce a correlation between consent and HIV status (which we censor for individuals with $Consent_i = 0$). As we know the true HIV prevalence, we can evaluate the relative performance of imputation, the standard selection model, and our copula selection model. By varying the structure of the error terms, we assess the extent to which the standard selection model is sensitive to the assumption of bivariate normality, and whether the copula approach can be used to correct for potential bias and inefficiency. We confirm that the imputation model performs poorly when there is correlation between consent and HIV status (bias of between 40% and 50%) and that selection models are appropriate for correcting for this correlation. We find that the performance of the bivariate normal selection model is related to the strength of the relation between the selection variable and consent in some simulation scenarios. This closely parallels the case of instrumental variables and is consistent with previous results (Leung and Yu, 1996). When the relation between interviewer identity and consent is less strong, bias and inefficiency can arise when the model is misspecified. For example, when normal errors are cubed, we find the mean bias of the standard Heckman–type model is 14%, whereas the bias in the copula model is less than half this amount; additionally, the copula model is more efficient. The distribution of the normal and copula estimators, along with that for the imputation model, is shown in Figure 2. Further details are presented in the eAppendix (http://links.lww.com/EDE/A858).

2.3 Data

We use data from the 2007 Zambia DHS (publicly available at www.dhsprogram.com). We adopt the same explanatory variables and specification as used in previous research (Hogan et al., 2012), the code for which is freely available online from http://hdl.handle.net/1902.1/17657. As outlined in model (1), interviewer identity enters into the consent equation as a series of dummy variables, one for each interviewer. As some interviewer fixed effects are collinear with other variables in the model, interviewers with fewer than 50 interviewees, or those with interviewer effects which are collinear, are combined into a single category. In the final selection models, there are 29 interviewers for men and 45 for women. We focus on estimating selection models for persons who refused to consent to an HIV test, as opposed to respondents who could not be contacted, because there are relatively few people who could not be contacted compared with people who refused to test. However, the methodology we propose could be easily applied to respondents who were not contacted. Table 3 illustrates the composition of the analysis sample for men and women separately. Excluding people who could not be contacted, of the eligible 6416 men, 1318 (21%) declined the offer of an HIV test; of the eligible 7025 women in the survey, 1400 (20%) declined the offer of an HIV test. Table 3 also illustrates the HIV prevalence estimates based on the complete case analyses (respondents with a valid HIV test), which is estimated to be 12% for men and 16% for women. All our estimates of HIV prevalence are weighted and take account of the complex survey design of the DHS (Corsi et al., 2012).

Table 3: Summary Statistics for Men and Women (Zambia Demographic and Health Survey 2007)

	HIV Prevalence		HIV Test	
	%	(95% CI)	Consented No. (%)	Refused No. (%)
Men	12	(11 to 13)	5,098 (79)	1,318 (21)
Women	16	(15 to 17)	5,625 (80)	1,400 (20)

Note: HIV prevalence estimates are based on analysis of respondents who have a valid HIV test and are adjusted for survey design. Noncontacts are excluded.

Statistical analyses were performed in R version 3.1.1 using the SemiParBIVProbit package (Marra and Radice, 2013b).

3 Results

Table 4 presents estimates for the rank association between consent to test and HIV status (Kendall's tau, τ) for each of the 9 copula models used, along with the corresponding 95% CI, which account for clustering at the primary sampling unit level. A measure of model fit (the AIC) is also presented in the final column of Table 4. Although the AIC is not adjusted for clustering, this limitation is unlikely to affect the preferred ordering of the models (Dziak and Li, 2006). For men, there is support for the hypothesis of selection bias, with a negative association for each of the copula models, and the 95% CI for τ excludes zero in each case. The τ of 0.53 for the normal model corresponds to a ρ (correlation coefficient) of 0.73. On the basis of the AIC, the model with the best fit is the C_{J90}. For women, the measure of association between testing and HIV status is also negative, although the association is less strong than for men, with the 95% CI in most models including zero. The τ of 0.19 in the normal model corresponds to a ρ of 0.30. On the basis of the AIC, the preferred copula specification for women is C_g or C_{C270}. Table 5 gives the corresponding HIV prevalence estimates. Point estimates for all copula models for men are similar, ranging from 19% to 21%, with the preferred model (C_{J90} copula) indicating a population HIV prevalence of 21% (with a corresponding 95% CI of 16% – 25%). As with men, HIV prevalence estimates for women are not sensitive to the choice of the copula function, ranging between 18% and 19%. The result for the preferred copula model (C_g) is 19% (with a 95% CI of 13% – 24%).

4 Discussion

Longitudinal evidence has demonstrated that people who do not consent to be tested in HIV surveys are more likely to be HIV positive than people who do consent to be tested (Bärnighausen et al., 2012; Floyd et al., 2013; Obare, 2010; Reniers and Eaton, 2009). Heckman–type selection models can be used to correct for the bias in data that are missing due to unobserved variables. However, the practical use of these selection models has been criticized for the strong assumptions required for their implementation (Arpino et al., 2014; Geneletti et al., 2011; Puhani, 2000). Our method provides estimates of HIV prevalence that are corrected for missing data on unobserved variables, without relying on the assumption of bivariate normality for identification. This study shows how the credibility of conclusions from selection models can be enhanced by demonstrating that the identification does not rely on a specific functional form for estimationhere, for the example of estimating HIV prevalence in Zambia. The wider variety of error distributions we consider

Table 4: Measures of Association Between Consent to be Tested for HIV and HIV Status for Men and Women (Zambia Demographic and Health Survey 2007)

	Men			Women		
Copula Model	Kendall's Tau	(95% CI)	AIC	Kendall's Tau	(95% CI)	AIC
Normal	-0.53	(-0.76 to -0.13)	9,672.52	-0.19	(-0.47 to 0.12)	11,237.27
Frank	-0.58	(-0.72 to -0.24)	9,667.97	-0.17	(-0.44 to 0.17)	11,237.54
Student T	-0.53	(-0.79 to -0.07)	9,675.19	-0.19	(-0.49 to 0.15)	11,238.36
Clayton 90	-0.31	(-0.8 to -0.05)	9,677.25	-0.13	(-0.6 to -0.02)	11,237.64
Clayton 270	-0.71	(-0.84 to -0.53)	9,666.31	-0.27	(-0.74 to -0.05)	11,237.27
Joe 90	-0.72	(-0.84 to -0.55)	9,666.21	-0.28	(-0.74 to -0.05)	11,237.28
Joe 270	-0.32	(-0.8 to -0.05)	9,678.22	-0.13	(-0.6 to -0.01)	11,237.89
Gumbel 90	-0.61	(-0.82 to -0.35)	9,670.97	-0.23	(-0.68 to -0.03)	11,237.37
Gumbel 270	-0.43	(-0.82 to -0.11)	9,676.32	-0.16	(-0.64 to -0.02)	11,237.69

Note: Estimates are presented for selection models based on the maximization of log-likehood (11) and the copula functions defined in Table 2. The AIC is shown in columns 3 and 5. The selection variable is a series of fixed effects for interviewer identity, of which there are 29 for men and 45 for women. Additional control variables include urban setting, region, interview language, ethnicity, religion, marital status, high-risk sexual behavior in the past year, condom use at last sex, sexually transmitted disease in the past year, tobacco and alcohol use, knowing someone with HIV/AIDS, willingness to care for a family member with HIV/AIDS, and having had a previous HIV test (Bärnighausen et al., 2011a; Hogan et al., 2012). Noncontacts are excluded. CI are adjusted for clustering at the primary sampling unit level.

Table 5: HIV Prevalence Estimates for Men and Women (Zambia Demographic and Health Survey 2007)

	Men		Women	
Copula Model	HIV Prevalence	(95% CI)	HIV Prevalence	(95% CI)
Normal	20	(13 to 28)	19	(13 to 24)
Frank	21	(15 to 26)	18	(14 to 23)
Student T	21	(13 to 29)	19	(14 to 25)
Clayton 90	19	(7 to 30)	19	(13 to 25)
Clayton 270	21	(16 to 25)	18	(14 to 22)
Joe 90	21	(16 to 25)	18	(14 to 22)
Joe 270	19	(8 to 31)	19	(12 to 26)
Gumbel 90	21	(14 to 27)	18	(14 to 23)
Gumbel 270	20	(10 to 30)	19	(13 to 25)

Note: HIV prevalence is based on individuals who have a valid HIV test and predicted HIV status from selection models based on the maximization of log-likehood (11) and the copula functions defined in Table 2. The selection variable is a series of fixed effects for interviewer identity, of which there are 29 for men and 45 for women. Additional control variables include urban setting, region, interview language, ethnicity, religion, marital status, high-risk sexual behavior in the past year, condom use at last sex, sexually transmitted disease in the past year, tobacco and alcohol use, knowing someone with HIV/AIDS, willingness to care for a family member with HIV/AIDS, and having had a previous HIV test (Bärnighausen et al., 2011a; Hogan et al., 2012). Noncontacts are excluded. CI are adjusted for clustering at the primary sampling unit level and prevalence estimates are weighted. The preferred model according to the AIC is Joe 90 for men and the Normal and Clayton 270 models for women.

provide a more meaningful assessment of the importance of the bivariate normality assumption than was previously possible using existing methods.

Our results indicate population HIV prevalence for men in the preferred selection model that is statistically larger than that based on the assumption of missing at random for the data on respondents who refuse to consent to be tested. The preferred copula model for men, the Joe 90 (C_{J90}), indicates the presence of tail dependence. This finding highlights the importance of our contribution of allowing for a large number of parametric structures. The previous literature relied on a more narrow set of models, which did not include the rotated Joe, Gumbel, or Clayton copulae (Radice et al., 2015). In addition, we find that the corresponding 95% CI for the Joe 90 copula estimate is substantially narrower than that obtained from the bivariate normal model, indicating an efficiency gain from implementing a dependence structure that may more accurately reflect the true underlying distribution of the data. In this analysis, imputation models, which require that the strong assumption of data being missing at random is met, produced results that are almost identical to the complete case analysis of respondents who have a valid HIV test, which is similar to previous findings (Hogan et al., 2012; Mishra et al., 2008; Zaidi et al., 2013). Given the increasing focus on treatment–as–prevention in HIV research and policy, it is likely that HIV surveys will increase in both frequency and coverage in many settings. Therefore, the issue of non–response bias in such surveys will likely increase in importance. Moreover, knowledge of HIV status, and therefore the potential for selection bias that depends on the unobserved variable HIV status, is also likely to increase as a result. The development of approaches to correct for selection on unobserved variables while relying on as few assumptions as possible, as well as approaches to test the robustness of the results from such selection models to variation in assumptions is important. The use of copula functions in Heckman–type selection models is a significant advance toward this aim. We believe that our approach using several parametric assumptions in the implementation of Heckman–type selection models makes the use of these models an even more viable alternative to the other approaches to correct for selection bias, which require the strong and untestable assumption that data are missing at random.

Our simulation results indicate that estimates obtained from the standard selection model that assumes bivariate normality can be biased and inefficient when the structure of the error term is misspecified. The copula models we propose perform well under a variety of different correlational structures, including scenarios with asymmetry. Although these conclusions are valid for the simulation settings considered here, it cannot be determined a priori whether relaxing the assumption of normality will lead to dramatically

different estimated prevalence as the error terms are not observed and the true structure is unknown. It is difficult to simulate the highly complex processes that likely underlie the relation between consent to HIV testing and HIV status. However, these results do suggest that there are a variety of scenarios where an incorrect normality assumption leads to biased results, and where the copula approach can correct for this bias.

The methodology we outline is easily implemented in standard statistical software (http://cran.r-project.org/web/packages/SemiParBIVProbit), and we provide the code for all the analyses discussed in this article (eAppendix, http://links.lww.com/EDE/A858, and online at http://dx.doi.org/10.7910/DVN/27727). Assessing the sensitivity of selection model results to relaxing the bivariate normality assumption is easily achieved with this approach, not only in the specific context of HIV prevalence estimation but also in other empirical applications.

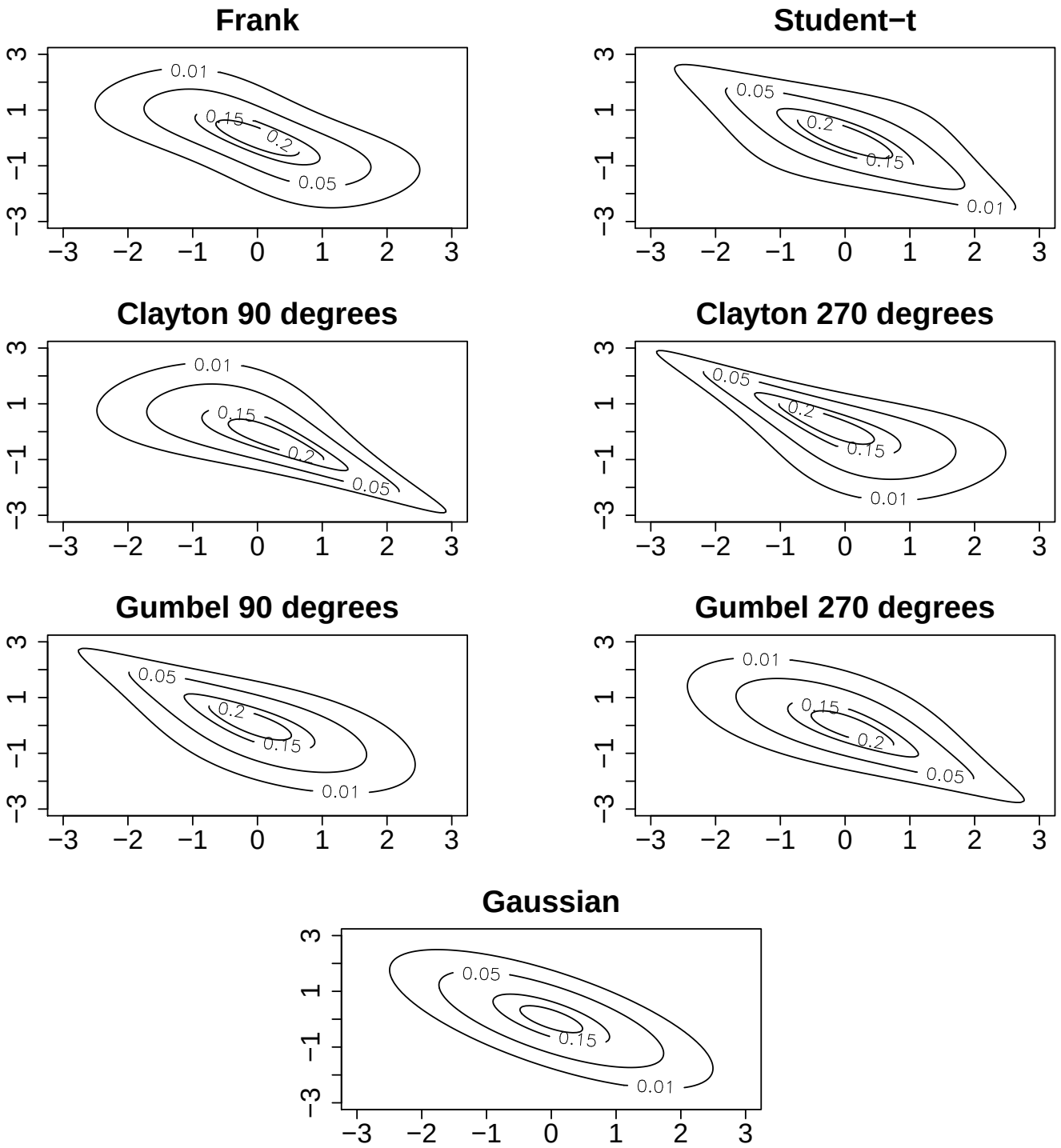
There are a number of avenues for future research. First, the literature on copula model selection for censored data is underdeveloped. Implementing goodness-of-fit tests is difficult due to the combination of censoring, the fact that the error terms are unobserved, and the fact that the outcomes are binary. We have focused here on conventional information criteria, but goodness- of-fit tests in this context are an important area for development, which could substantially improve the performance of copula models. Second, there are advantages and disadvantages associated with the copula approach compared with semiparametric and nonparametric models. The latter have the advantage of not requiring the true parametric model to be specified by the researcher. However, although theoretically possible (De Luca, 2008), the intercept is typically not identified in these models, and so this approach is not suitable for estimating population means based on binary outcomes, such as HIV prevalence. Semiparametric approaches that allow for the estimation of the intercept require additional assumptions and have only been developed for the case of continuous outcomes (Andrews and Schafgans, 1998; Schafgans and Zinde-Walsh, 2002). Additionally, the semiparametric approaches typically generate estimates that are inefficient relative to fully parameterized models, may not allow diagnostics, are limited with regard to the inclusion of a large set of covariates, and may be computationally demanding (Bhat and Eluru, 2009). In contrast, the computational simplicity of the copula approach allows the practitioner to exploit familiar tools such as maximum likelihood without requiring simulation methods or numerical integration. Maximum likelihood, in turn, allows for the simultaneous estimation of all the parameters of the model and, if the usual regularity conditions are met, ensures consistent, efficient, and asymptotically normal estimators (Smith, 2003). Finally, copula modelling allows for direct estimation

of the dependence structure in the sample selection model, whereas semiparametric methods do not (Genius and Strazzera, 2008).

Further analysis should focus on establishing the validity of the other main requirement in sample selection models underlying the estimation of HIV prevalence in the presence of nonresponse, namely the existence of a selection variable that does not independently affect the outcome of interest. The recording of additional information on interviewer characteristics in HIV surveys, such as their age, sex, and experience would facilitate such research (Clark and Houle, 2014). Although interviewer identity is plausibly a function only of survey design, and not related to individual-level characteristics, this claim is difficult to prove conclusively. As a robustness check, we included a cluster random effect in our model using a 2-stage procedure to account for potential correlation between interviewer allocation and the characteristics of the individual's primary sampling unit (McGovern et al., 2015). HIV prevalence estimates in this analysis were similar, but this approach is inefficient and resulted in an attenuated relation between consent and interviewer identity. Therefore, incorporating random effects directly into these types of selection models is another important direction for future research. Random effects could potentially be included along with a flexible semiparametric approach to modelling covariates (Marra and Radice, 2013a). In general, as we never observe the HIV status of respondents who refuse to be tested, establishing whether estimates based on selection models can be supported with objective external data, such as alternative selection variables or mortality records, would help validate this approach.

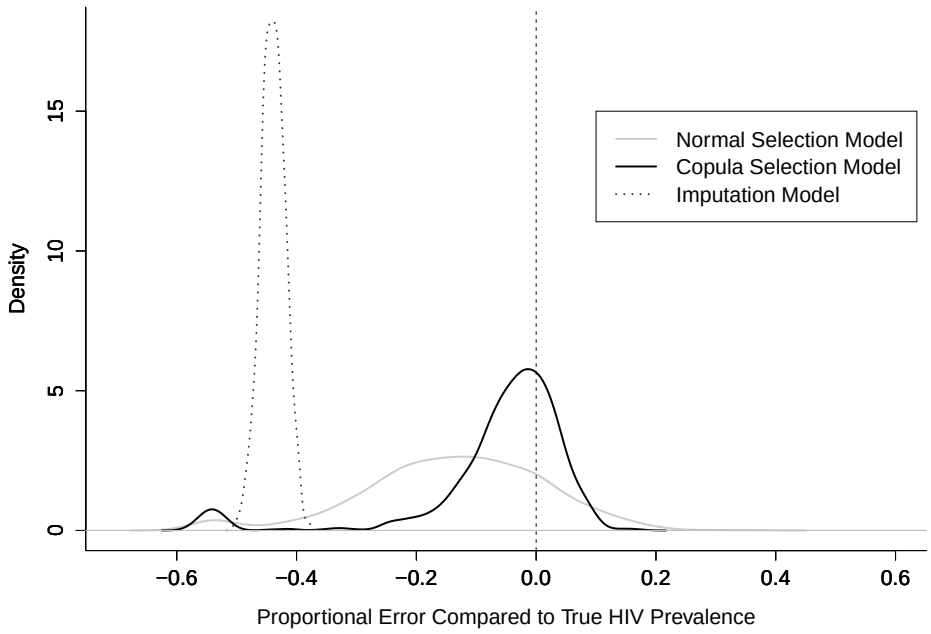
In sum, we introduce and demonstrate a new approach for relaxing the assumption of bivariate normality in Heckman–type selection models with binary outcomes using copulae. Our simulation study illustrates that this methodology can be used to correct for the bias and inefficiency associated with misspecification of the dependence structure between selection into the data and the outcome of interest. In empirical work, establishing that selection model estimates are robust to alternative functional form specifications for the relation between selection and the outcome increases the credibility of these estimates.

Figure 1: Illustration of Modelling Dependence using Copulae



Note: Observations are drawn from the corresponding bivariate distributions with $n = 1,000$ and $\tau = 0.50$. We do not show the Joe 90 and Joe 270 copulae, these are similar in appearance to the corresponding Gumbel versions.

Figure 2: Simulation Results for HIV Prevalence Estimates with Non-normal Errors



Note: This scenario illustrates the case with cubed normal errors. The distribution of the proportional error of estimates of HIV prevalence obtained from the normal selection model (Gaussian copula), a copula selection model and an imputation model are shown. The simulation is based on the 2007 Zambia Demographic and Health Survey for men, with $n = 6,500$ and 1,000 replications. For each replication, the proportional error for each estimator is calculated as $mean\frac{(HIVModel-HIVTrue)}{HIVTrue}$. The copula model is defined as the copula with the best fit in each replication according to the AIC. Errors for the latent variables for consent and HIV status were drawn from a bivariate normal distribution with $mean = 0$ and $\tau = 0.50$, cubed, and then scaled to have mean 0. The mean true HIV prevalence was 21%, observed HIV prevalence (for those with $Consent = 1$) was 12%. Consent to be tested was 81%, and the F statistic for interviewer identity was 3.5. The F statistic is calculated as a joint test of significance for interviewer identity in a regression of consent on interviewer identity with the inclusion of the model control variables. See the eAppendix (http://links.lww.com/EDE/A858, and online at http://dx.doi.org/10.7910/DVN/27727) for further details, including the R code for replicating the simulations.

References

D. W. Andrews and M. M. Schafgans. Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies*, 65(3):497–517, 1998.

B. Arpino, E. D. Cao, and F. Peracchi. Using panel data for partial identification of human immunodeficiency virus prevalence when infection status is missing not at random. *Journal of the Royal Statistical Society: Series A*, 177(3):587–606, 2014.

T. Bärnighausen, J. Bor, S. Wandira-Kazibwe, and D. Canning. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology*, 22(1):27–35, 2011a.

T. Bärnighausen, J. Bor, S. Wandira-Kazibwe, and D. Canning. Interviewer identity as exclusion restriction in epidemiology. *Epidemiology*, 22(3):446, 2011b.

T. Bärnighausen, F. Tanser, A. Malaza, K. Herbst, and M. Newell. HIV status and participation in HIV surveillance in the era of antiretroviral treatment: a study of linked population-based and clinical data in rural South Africa. *Tropical Medicine & International Health*, 17(8):e103–e110, 2012.

C. R. Bhat and N. Eluru. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological*, 43(7):749–765, 2009.

J. T. Boerma, P. D. Ghys, and N. Walker. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *The Lancet*, 362(9399):1929–1931, 2003.

E. C. Brechmann and U. Schepsmeier. Modeling dependence with C-and D-vine copulas: The R-package CDVine. *Journal of Statistical Software*, 52:1–27, 2012.

S. Clark and B. Houle. Evaluation of Heckman Selection Model Method for Correcting Estimates of HIV Prevalence from Sample Surveys via Realistic Simulation. *Center for Statistics and the Social Sciences Working Paper No. 120, University of Washington*, 2012.

S. J. Clark and B. Houle. Validation, Replication, and Sensitivity Testing of Heckman-Type Selection Models to Adjust Estimates of HIV Prevalence. *PloS one*, 9(11):e112563, 2014.

D. Conniffe and D. O'Neill. Efficient Probit Estimation with Partially Missing Covariates. *Advances in Econometrics*, 27:209–245, 2011.

D. J. Corsi, M. Neuman, J. E. Finlay, and S. Subramanian. Demographic and Health Surveys: a profile. *International Journal of Epidemiology*, 41(6):1602–1613, 2012.

D. Dancer, A. Rammohan, and M. D. Smith. Infant mortality and child nutrition in Bangladesh. *Health Economics*, 17(9):1015–1035, 2008.

G. De Luca. SNP and SML estimation of univariate and bivariate binary-choice models. *Stata Journal*, 8 (2):190–220, 2008.

A. R. T. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, 2006.

J. A. Dubin and D. Rivers. Selection bias in linear regression, logit and probit models. *Sociological Methods & Research*, 18(2-3):360–390, 1989.

J. Dziak and R. Li. Variable Selection with Penalized Generalized Estimating Equations. *The Methodology Center, Pennsylvania State University*, (Technical Report 06-78), 2006.

S. Floyd, A. Molesworth, A. Dube, A. C. Crampin, R. Houben, M. Chihana, A. Price, N. Kayuni, J. Saul, and N. French. Underestimation of HIV prevalence in surveys when some people already know their status, and ways to reduce the bias. *AIDS*, 27(2):233–242, 2013.

S. Geneletti, A. Mason, and N. Best. Adjusting for selection effects in epidemiologic studies: why sensitivity analysis is the only solution? *Epidemiology*, 22(1):36–39, 2011.

M. Genius and E. Strazzera. Applying the copula approach to sample selection modelling. *Applied Economics*, 40(11):1443–1455, 2008.

M. Gersovitz. HIV testing: principles and practice. *World Bank Research Observer*, 26(1):1–41, 2011.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

D. R. Hogan, J. A. Salomon, D. Canning, J. K. Hammitt, A. M. Zaslavsky, and T. Bärnighausen. National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models. *Sexually transmitted infections*, 88(Suppl 2):i17–i23, 2012.

W. Janssens, J. van der Gaag, T. F. R. de Wit, and Z. Tanovia. Refusal bias in the estimation of HIV prevalence. *Demography*, 51(3):1131–1157, 2014.

K. Kranzer, N. McGrath, J. Saul, A. C. Crampin, A. Jahn, S. Malema, D. Mulawa, P. E. Fine, B. Zaba, and J. R. Glynn. Individual, household and community factors associated with HIV test refusal in rural Malawi. *Tropical Medicine & International Health*, 13(11):1341–1350, 2008.

S. F. Leung and S. Yu. On the choice between sample selection and two-part models. *Journal of Econometrics*, 72(1):197–229, 1996.

T. Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19, 2004.

D. Madden. Sample selection versus two-part models revisited: The case of female smoking and drinking. *Journal of Health Economics*, 27(2):300–307, 2008.

G. Marra and R. Radice. A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electronic Journal of Statistics*, 7:1432–1455, 2013a.

G. Marra and R. Radice. SemiParBIVProbit: Semiparametric Bivariate Probit Modelling. *R package version 3.2-11*, 2013b.

M. Marston, K. Harriss, and E. Slaymaker. Non-response bias in estimates of HIV prevalence due to the mobility of absentees in national population-based surveys: a study of nine national surveys. *Sexually Transmitted Infections*, 84(Suppl 1):i71–i77, 2008.

M. McGovern, T. Bärnighausen, J. Salomon, and D. Canning. Using Interviewer Random Effects to Calculate Unbiased HIV Prevalence Estimates in the Presence of Non-Response: a Bayesian Approach. *BMC Medical Research Methodology*, 15(8), 2015.

V. Mishra, B. Barrere, R. Hong, and S. Khan. Evaluation of bias in HIV seroprevalence estimates from national household surveys. *Sexually Transmitted Infections*, 84(Suppl 1):i63–i70, 2008.

J. M. Murteira and Ó. D. Lourenço. Health care utilization and self-assessed health: specification of bivariate models using copulas. *Empirical Economics*, 41(2):447–472, 2011.

F. Obare. Nonresponse in repeat population-based voluntary counseling and testing for HIV in rural Malawi. *Demography*, 47(3):651–665, 2010.

C. O'Muircheartaigh and P. Campanelli. The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A*, 161(1):63–77, 1998.

J. E. Prieger. A flexible parametric selection model for non?normal data with application to health care usage. *Journal of Applied Econometrics*, 17(4):367–392, 2002.

P. Puhani. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14 (1):53–68, 2000.

R. Radice, G. Marra, and M. Wojtys. Copula Regression Spline Models for Binary Outcomes. *Statistics and Computing*, Forthcoming, 2015.

G. Reniers and J. Eaton. Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. *AIDS*, 23(5):621–629, 2009.

G. Reniers, T. Araya, Y. Berhane, G. Davey, and E. J. Sanders. Implications of the HIV testing protocol for refusal bias in seroprevalence surveys. *BMC Public Health*, 9(1):1–9, 2009.

M. Schafgans and V. Zinde-Walsh. On intercept estimation in the sample selection model. *Econometric Theory*, 18(01):40–50, 2002.

M. D. Smith. Modelling sample selection using Archimedean copulas. *The Econometrics Journal*, 6(1): 99–123, 2003.

O. Sterck. Why Are Testing Rates So Low in Sub-Saharan Africa? Misconceptions and Strategic Behaviors. In *Forum for Health Economics and Policy*, volume 16, pages 219–257, 2013.

P. K. Trivedi and D. M. Zimmer. Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111, 2007.

W. P. Van de Ven and B. Van Praag. The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics*, 17(2):229–252, 1981.

F. Vella. Estimating models with sample selection bias: a survey. *Journal of Human Resources*, 33(1): 127–169, 1998.

R. Winkelmann. Copula Bivariate Probit Models: With an Application to Medical Expenditures. *Health Economics*, 21(12):1444–1455, 2012.

J. Zaidi, E. Grapsa, F. Tanser, M.-L. Newell, and T. Bärnighausen. Dramatic increase in HIV prevalence after scale-up of antiretroviral treatment. *AIDS*, 27(14):2301–2305, 2013.

Simulation Details and R Code Appendix For:

On the Assumption of Bivariate Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence*

Mark E. McGovern[†1], Till Bärnighausen[2,3], Giampiero Marra[4], and Rosalba Radice[5]

[1]Harvard University
[2]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, USA
[3]Wellcome Trust Africa Centre for Health and Population Studies, University of KwaZulu-Natal, South Africa
[4]Department of Statistical Science, University College London
[5]Department of Economics, Mathematics and Statistics, Birkbeck, University of London

March 2015

Abstract

This Appendix describes our simulation study for evaluating the performance of copula based selection models for binary outcomes in further detail. We outline the procedure for generating the simulated data, and present the results. We construct latent variables for consent to test for HIV and HIV status, which incorporate four different dependence structures (bivariate normal (equivalent to the Gaussian copula), Student-t copula, bivariate normal cubed, and Clayton 270 copula). For each of these four dependence structures, we consider two cases, one with a weak association between interviewer identity and consent, and one with a stronger association between interviewer identity and consent, giving 8 simulation scenarios in total. We also provide the relevant R code for replicating the simulation study.

JEL Classification: *C34, J11*

Keywords: *Selection Bias, Bivariate Normality, Copula, HIV*

*Published as: McGovern, M.E., Bärnighausen, T., Marra, G., Radice, R., 2015. On the Assumption of Bivariate Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence. Epidemiology 26, 229 – 327. Available at http://dx.doi.org/10.1097/EDE.0000000000000218. We are grateful to Slawa Rokicki for invaluable suggestions and advice. We also thank the editor, two anonymous referees, seminar participants at Harvard University, University College London, and the UNAIDS Reference Group on Estimates, Modelling, and Projections for comments. Mark McGovern received financial support from the Program on the Global Demography of Aging which receives funding from the National Institute on Aging (grant no. 1 P30 AG 024409-09). Till Bärnighausen received financial support through grant 1 R01 HD 058482 01 from the National Institute of Child Health and Human Development, National Institutes of Health.

†Corresponding author. Current Address: Queen's University Belfast, Riddel Hall, 185 Stranmillis Road, Belfast, BT9 5EE, Northern Ireland. Email: markemcgovern@gmail.com.

Contents

1 Details of the Simulation Study

Outline

We simulate an HIV survey with missing data in which the assumption of missing at random does not hold. We follow the approach implemented in Clark and Houle (2012) by generating a dataset based on a real HIV survey, in this case the 2007 Zambia Demographic and Health Survey (DHS) for men.[1] Therefore, our simulations closely match the overall observed consent rates and HIV prevalence in the actual data used in the empirical part of this paper (HIV prevalence of 12% and a consent rate of 79%), although unlike Clark and Houle (2012) we do not attempt to match covariate specific HIV prevalence rates. For each individual in the simulated dataset, we construct latent variables for consent and HIV status based on two observed covariates (age and urban or rural place of residence). We use place of residence as our second covariate rather than sex, as all our empirical models are stratified by sex and thus could not be included as a covariate. The distributions of the two observed covariates are drawn to match those in the data, see table e1 for a description of these characteristics.

Following our empirical model outlined in equations e1-e4 and described in further detail in the main text, observed consent and HIV status are based on latent variables for consent and HIV status for individual i with interviewer j,[2;3;4] which are determined by the observed covariates X_{ij}, interviewer identity Z_j, and corresponding error terms u_{ij} and ϵ_{ij} in both equations:

$$Consent^*_{ij} = X^T_{ij}\beta + Z^T_j\alpha + u_{ij},\ i = 1...n,\ j = 1...J \tag{A1}$$

$$Consent_{ij} = 1 \text{ if Consent}^*_{ij} > 0,\ \text{Consent}_{ij} = 0 \text{ otherwise,} \tag{A2}$$

$$HIV^*_{ij} = X^T_{ij}\gamma + \epsilon_{ij}, \tag{A3}$$

$$HIV_{ij} = 1 \text{ if HIV}^*_{ij} > 0,\ \text{HIV}_{ij} = 0 \text{ otherwise.} \tag{A4}$$

Table A1: Summary Statistics for Simulated Data

Age Category	N	%	Urban/Rural Place of Residence	N	%
15-19	1,367	21	Urban	2,820	43
20-24	1,074	17	Rural	3,680	57
25-29	1,044	16	Total	6,500	100
30-34	904	14			
35-39	711	11			
40-44	486	7			
45-49	418	6			
50-54	271	4			
55-59	225	3			
Total	6,500	100			

T denotes the transpose function. Individuals are matched to one of 30 interviewers (Z_j), whose persuasiveness (λ_j) is drawn from a standard normal distribution: $\lambda \sim N(0,1)$. Interviewer persuasiveness is included in the latent consent equation ($Consent^*_{ij}$), but excluded from the latent HIV equation (HIV^*_{ij}). In real data this must be assumed as it is not generally possible to test this condition without additional external information, but here we impose that the selection variable is valid (i.e. the exclusion restriction holds). We construct the latent consent equation using persuasiveness, but we estimate the selection models using interviewer identity as the selection variable, because this is the information we observe in practice in the data. Both the latent consent and latent HIV status equations (shown in equations e5 and e6) are constructed from their corresponding linear predictors which are given by the linear combinations of the covariates (age category and place of residence) and the regression parameter vectors (β, γ) estimated by fitting a bivariate sample selection model on the 2007 Zambia DHS for men. These parameter values are summarized in table e2. The corresponding error terms, u_{ij} and ϵ_{ij}, the joint distribution of which we vary according to the scenario of interest (bivariate normality or the alternatives), are also added to the linear predictors.

$$Consent^*_{ij} = \lambda_j\delta + \beta_1 + \sum_{k=2}^{8}\beta_k I[AgeGroup_{kij}] + \beta_9 I[Rural_{ij}] + u_{ij}, \tag{A5}$$

$$HIV^*_{ij} = \gamma_1 + \sum_{k=2}^{8}\gamma_k I[AgeGroup_{kij}] + \gamma_9 I[Rural_{ij}] + \epsilon_{ij} \tag{A6}$$

$I[\bullet]$ is the indicator variable, taking the value one if the individual is a member of age group k and 0 otherwise (for example). First we draw the jointly distributed error terms; then we calculate the linear predictors and hence the latent variables; finally, we generate the main binary outcomes of interest (consent and HIV status) which take the value 1 if the latent variable is > 0, and 0 if the latent variable is ≤ 0.

In the simulated data we observe HIV status for everyone; however, in practice we only observe the HIV status of those who consent to test:

$$\text{HIV}_{ij} \text{ observed only if Consent}_{ij} = 1, \text{ missing otherwise.} \tag{A7}$$

Therefore, for our comparison of the performance of the analytic models, we censor the HIV outcome for individuals with consent=0. The structure of the error terms (specifically the joint distribution of u_{ij} and ϵ_{ij}), will determine the direction and extent of selection bias. This allows us to compare the true HIV prevalence (which we know) to that which would actually be observed in practice when there is missing data for HIV status because of refusal to test (or other mechanisms for missing data). We compare the result obtained from the selection and imputation models to the known true value. By varying the structure of the error terms, we can evaluate the extent to which the standard selection model is sensitive to the assumption of bivariate normality, and whether the copula approach can be used to correct for any potential bias and inefficiency.

The regression parameters used to generate the latent consent and HIV status variables in the simulated data are those which are observed in the bivariate model for consent and HIV status in the actual Zambian data, and do not vary across simulation scenarios. The two parameters which do vary are those which determine the structure of the error terms, and the strength of the association between interviewer identity and consent (δ).

Table A2: Summary of Simulation Parameters

Sample Size	6,500
Number of Interviewers	30
Number of Replications	1,000
Kendall's Tau (τ)	-0.50

Regression Parameters for Latent Dependent Variables

	Consent Equation	HIV Status Equation
Interviewer Persuasiveness (δ)	0.25 or 0.50	
Control Variable (X_{ij})	β_k	γ_k
Constant	0.604	-1.156
Age 15-19	(Omitted Category)	(Omitted Category)
Age 20-24	-0.039	0.229
Age 25-29	-0.036	0.703
Age 30-34	0.017	1.036
Age 35-39	0.081	1.147
Age 40-44	0.134	1.203
Age 45-49	0.053	1.063
Age 50-54	0.028	0.834
Age 55-59	0.166	0.661
Rural	0.123	-0.396

HIV prevalence and consent will depend to a certain extent on error structure as we construct these outcomes from the latent variables. We allow the persuasiveness of interviewers to vary, as the exclusion restriction in sample selection models closely parallels that in instrumental variables (IV) analysis,[5] where the strength of the instrumental variable (defined as we explain below) has an important impact on the performance of the model.[6] While in theory we could vary other aspects of the simulated data (such as sample size, the amount of missing data, HIV prevalence, the degree of selection bias measured by the association between the error terms, the inclusion of additional covariates, or violation of the exclusion restriction), in practice the model is already complex, and the true data generating process is unknown, therefore we opt for this more parsimonious approach. See Clark and Houle (2012) for further discussion of these factors.[1]

We consider four different scenarios for the error terms; bivariate normal (equivalent to the Gaussian copula), a Student-t copula, a case where we cube the bivariate normal errors as in Clark and Houle (2012),[1] and a Clayton 270 copula. In each case, errors were constructed to have mean 0, and a Kendall's Tau, (τ, a nonparametric of association between u_{ij} and ϵ_{ij}) of -0.50, which is the value we obtained in our empirical analysis of the actual data. In the scenario with normal cubed errors, u_{ij} and ϵ_{ij} were drawn from a bivariate normal distribution and then cubed. For each of the four error scenarios, we also report two cases, one with weaker interviewer persuasiveness ($\delta = 0.25$), and one with stronger interviewer persuasiveness ($\delta = 0.50$), giving eight simulation scenarios in total (summarised in table e3). When we increased the interviewer effect further, we found that the performance of the standard Heckman selection model based on bivariate normality generally improved (although we only considered scenarios with the sample characteristics, such

as sample size, consent rates etc., held constant). When the relationship between interviewer identity and consent was weak, we found that bias can arise when the model is misspecified. Therefore, in these results we focus on reporting the case where the F statistic was close to 10 (with is a value commonly used a rule of thumb for having a weak instrument in IV analysis).[6]

Table A3: Summary of the Eight Simulation Scenarios

Error Term Structure (u_{ij}, ϵ_{ij})	Strength of Interviewer Persuasiveness (δ)	
	Scenario 1	Scenario 2
Gaussian Copula (Bivariate Normal)	Weak ($\delta = 0.25$)	Stronger ($\delta = 0.50$)
Student-t Copula	Weak ($\delta = 0.25$)	Stronger ($\delta = 0.50$)
Bivariate Normal Cubed	Weak ($\delta = 0.25$)	Stronger ($\delta = 0.50$)
Clayton 270 Copula	Weak ($\delta = 0.25$)	Stronger ($\delta = 0.50$)

For each of the eight simulation scenarios, we compare the mean proportional error, calculated as $mean\frac{(HIV_{Model}-HIV_{True})}{HIV_{True}}$, for each of the following models: Gaussian (C_g), which is equivalent to the standard bivariate normal probit model; Frank (C_f); 90 and 270 degrees rotated Clayton ($C_{c_{90}}$,$C_{c_{270}}$); 90 and 270 degrees rotated Joe ($C_{J_{90}}$,$C_{J_{270}}$); 90 and 270 degrees rotated Gumbel ($C_{G_{90}}$,$C_{G_{270}}$); Student-t (C_t), and an imputation-based estimate using the MICE package.[7] We also considered more complex imputation approaches and found similar results. We also consider the root mean square error (RMSE), calculated as $\sqrt{mean(HIV_{Model} - HIV_{True})^2}$, and calculate the partial F statistic for interviewer identity in the consent equation, which we obtain by implementing a linear probability model for consent on the covariates (age category and place of residence) and interviewer identity, and running a joint test of statistical significance on the interviewer identity parameters. All analyses were conducted in R, version 3.1,[8] using the SemiParBIVProbit package.[9] The following section gives the relevant R code. See the main text for a graphical representation of these copulae and further details on the implementation of these models. We also use the following other R packages in the analysis and to generate this document: aod,[10] CDVine,[11] copula,[12] doParallel,[13] doRNG,[14] foreach,[15] and knitr.[16]

Results

We summarize our findings in table e4, which compares the results from the standard Heckman-type selection model (using the assumption of bivariate normality), the imputation model, and the preferred copula model. For the scenarios with normal errors and normal cubed errors, the preferred copula model is the copula with the best fit, as determined by the AIC (Akaike Information Criterion). For the other scenarios (Student-t copula, Clayton 270 copula), the preferred copula for comparison is the underlying true model. In the first two error scenarios (normal errors and Student-t errors) the standard Heckman selection model performs well, however bias emerges in the next two error scenarios (normal cubed errors and Clayton 270 errors). For example, when the normal errors are cubed and the F statistic is around 3.5 we find that the mean bias of the standard Heckman model is -14%. In contrast, our copula model gives a mean bias of -6% in this scenario, and the associated RMSE is also lower at 0.029 compared to 0.042 for the standard bivariate normal model. The sampling distribution of the two estimators in this case, along with the imputation model, is shown in figure 2 in the main text. In addition, the mean bias for the bivariate normal model in the case with Clayton 270 errors and an F statistic of 6.8 is 12% compared to -2.7% for the copula model. Likewise

the RMSE is also larger for the normal model in this case (0.039 compared to 0 029). It is important to note that the imputation model performs poorly in all cases (bias of 40% to 50%).

While some minimal bias may remain in some cases in the copula models, this simulation analysis demonstrates that they are still preferable to the default normal model in the cases we examined. For example, a mean bias of -6% for the copula model in the scenario with normal cubed errors corresponds to a mean HIV prevalence estimate of 19.6%, compared to true mean prevalence of 20.8%, mean observed HIV prevalence of 11.7%, mean imputed HIV prevalence of 11.9%, and a mean bivariate normal selection model HIV prevalence estimate of 17.9%. The bias of the copula model is thus less than half that of the standard Heckman model. In addition, the copula model is more efficient. However, there are important directions for future research. As in Clark and Houle (2012),[1] we find some scenarios can still result in convergence failures. Here, we exclude them from the analysis. Work is ongoing to improve the implementation of these models. We only considered a limited number of scenarios, and factors such as sample size and amount of missing data could have an impact on the relative performance of copula models. Similarly, alternative specifications with additional covariates and potential interactions could also affect results, but this is difficult to assess given our limited knowledge about the true data generating process. In practice, violation of the exclusion restriction (which would occur if interviewer identity was associated with HIV status) is likely to impact on the model performance,[17] however we only considered cases where this assumption was valid by construction. Finally, we choose the preferred copula model on the basis of conventional information criteria. However, the literature on copula model selection for censored data is underdeveloped, and the implementation of more appropriate goodness of fit tests could substantially improve the performance of copula models in future applications.

The simulation procedure is based on the 2007 Zambia Demographic and Health Survey for men, with $N = 6,500$ and 1,000 replications. The first panel shows the true HIV prevalence, the observed HIV prevalence (defined as HIV prevalence among those with $Consent = 1$), the consent rate, and the partial F statistic for interviewer identity in the consent equation. Each of these is the mean value from the simulations. The mean proportional error of estimates of HIV prevalence obtained from the normal selection model (Gaussian Copula), a Copula selection model and an imputation model are shown in panel two. In the scenarios with normal, Student-t, and Clayton 270 errors, errors were drawn from the relevant bivariate distributions with $\tau = -0.50$ and scaled to have mean= 0. For the scenario with normal cubed errors, errors were drawn from a bivariate normal distribution with $\tau = -0.50$ and then cubed. For each replication, the proportional error for each estimator was calculated as $mean\frac{(HIV_{Model} - HIV_{True})}{HIV_{True}}$. In the scenarios with normal errors and normal cubed errors the copula model is defined as the copula with the best fit in each replication according to the Akaike Information Criterion. In the scenarios with Clayton errors and Student-t errors, the copula model is defined as the true copula model. As well as the standard bivariate normal probit model (Gaussian copula, C_g), the other copulae estimates are the Frank (C_f); 90 and 270 degrees rotated Clayton (C_{c90},C_{c270}); 90 and 270 degrees rotated Joe (C_{J90},C_{J270}); 90 and 270 degrees rotated Gumbel (C_{G90},C_{G270}); and Student-t (C_t). The F statistic is calculated as a joint test of significance for interviewer identity in a regression of consent on interviewer identity with the inclusion of the model control variables. The third panel shows the Root Mean Square Error (RMSE), which is defined as $\sqrt{mean(HIV_{Model} - HIV_{True})^2}$. The R code for implementing these models is given in following section.

Table A4: Comparison of Bivariate Normal and Copula Sample Selection Models

Type of Error Structure	True HIV	Selected HIV	Consent	Interviewer F
Normal ($\delta = 0.25$)	0.235	0.134	0.750	8.407
Normal ($\delta = 0.50$)	0.235	0.139	0.733	30.200
Student-t ($\delta = 0.25$)	0.265	0.142	0.701	6.751
Student-t ($\delta = 0.50$)	0.265	0.152	0.700	24.978
Normal Cubed ($\delta = 0.25$)	0.208	0.117	0.805	3.525
Normal Cubed ($\delta = 0.50$)	0.208	0.115	0.763	32.213
Clayton-270 ($\delta = 0.25$)	0.265	0.136	0.702	6.716
Clayton-270 ($\delta = 0.50$)	0.265	0.146	0.700	24.905

Mean Proportional Error

Type of Error Structure	Imputation Model	Standard (Gaussian) Selection Model	Copula Selection Model
Normal ($\delta = 0.25$)	-0.434	-0.002	0.010
Normal ($\delta = 0.50$)	-0.412	-0.002	0.001
Student-t ($\delta = 0.25$)	-0.470	-0.018	0.015
Student-t ($\delta = 0.50$)	-0.430	-0.026	0.006
Normal Cubed ($\delta = 0.25$)	-0.439	-0.138	-0.060
Normal Cubed ($\delta = 0.50$)	-0.447	-0.052	-0.085
Clayton-270 ($\delta = 0.25$)	-0.491	0.115	-0.027
Clayton-270 ($\delta = 0.50$)	-0.452	0.091	-0.007

Root Mean Square Error

Type of Error Structure	Imputation Model	Standard (Gaussian) Selection Model	Copula Selection Model
Normal ($\delta = 0.25$)	0.102	0.021	0.024
Normal ($\delta = 0.50$)	0.097	0.011	0.015
Student-t ($\delta = 0.25$)	0.125	0.031	0.030
Student-t ($\delta = 0.50$)	0.114	0.016	0.014
Normal Cubed ($\delta = 0.25$)	0.091	0.042	0.029
Normal Cubed ($\delta = 0.50$)	0.093	0.019	0.022
Clayton-270 ($\delta = 0.25$)	0.130	0.039	0.029
Clayton-270 ($\delta = 0.50$)	0.121	0.028	0.013

Appendix References

[1] S.J. Clark and B. Houle. Evaluation of Heckman Selection Model Method for Correcting Estimates of HIV Prevalence from Sample Surveys via Realistic Simulation. *Center for Statistics and the Social Sciences Working Paper No. 120, University of Washington*, 2012.

[2] Jeffrey A Dubin and Douglas Rivers. Selection bias in linear regression, logit and probit models. *Sociological Methods & Research*, 18(2-3):360–390, 1989.

[3] James J Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[4] Wynand PMM Van de Ven and Bernard Van Praag. The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics*, 17(2):229–252, 1981.

[5] Edward Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341, 2002.

[6] James H Stock, Jonathan H Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4), 2002.

[7] Stef Buuren and Karin Groothuis-Oudshoorn. MICE: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45(3), 2011.

[8] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Technical report, 2014.

[9] Giampiero Marra and Rosalba Radice. SemiParBIVProbit: Semiparametric bivariate probit modelling. *R package version 3.2-12*, 2014.

[10] Matthieu Lesnoff and Renaud Lancelot. Aod: analysis of overdispersed data. *R package version*, 1, 2012.

[11] Eike Christian Brechmann and Ulf Schepsmeier. Modeling dependence with C-and D-vine copulas: The R-package CDVine. *Journal of Statistical Software*, 52:1–27, 2012.

[12] Jun Yan. Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, 21(4):1–21, 2007.

[13] Revolution Analytics and Steve Weston. doparallel: Foreach parallel adaptor for the parallel package. *R package version*, 1(8), 2014.

[14] Renaud Gaujoux. dorng: Generic reproducible parallel backend for foreach loops. *R package version*, 1 (6), 2014.

[15] Revolution Analytics. foreach: Foreach looping construct for r. *R package version*, 1, 2013.

[16] Yihui Xie. knitr: A general-purpose package for dynamic report generation in r. *R package version*, 1 (7), 2013.

[17] David Madden. Sample selection versus two-part models revisited: The case of female smoking and drinking. *Journal of Health Economics*, 27(2):300–307, 2008.

2 Simulation R Code

2.1 File 1: 1_simulation.R

```
################################################################################################

# October 2014

# Simulation Code For:

# On the Assumption of Bivariate Normality in Selection Models:
# A Copula Approach Applied to Estimating HIV Prevalence

# Paramaterisation is based on the 2007 Zambia Demographic and Health Survey (Men)
# Data are publically available from www.dhsprogram.com

# The simulation setup for the Heckman model is based on Clark and Houle (2012)

# Copula models are implemented using the SemiParBIVProbit R package

# Use foreach package for multicore support and faster looping

# Results may differ slightly depending on factors such as the version of the package
# used or the number of computer cores

# We evaluate prevalence estimates from complete case analysis
# (cases with a valid HIV test), imputation, bivariate normal
# and copula based models

# See the main text and above for further details

# Four error scenarios: bivariate normal, Student-t copula, normal cubed, Clayton 270
# For each of these error scenarios there are two scenarios based on weaker and
# stronger interviewer effects, making 8 scenarios in total

# The following R packages are required: SemiParBIVProbit, copula, CDVine, foreach
# mice, aod, stats, doRNG, doParallel

# 10 files required: 1_simulation.R, 2_iteration.R, 3_best_fit.R, 4_parameters.R,
# 5_results.R, 6_tables.R, 7_normal_errors.R, 8_student_errors.R,
# 9_normal_cubed_errors.R, 10_clayton_errors.R

# Their contents can be copied from the text below

# These will need to be saved in the working directory

################################################################################################

# File 1/10
# Filename: 1_simulation.R
```

```r
# Main R file for implementing the simulation

# Clear files and Set WD

rm(list=ls())

# Load librarys
library(SemiParBIVProbit)
library(copula)
library(CDVine)
library(foreach)
library(mice)
library(aod)
require(stats)
library(doRNG)
library(doParallel)

setwd("C:/Users/User/Desktop/R")

# Optionally set number of processors for multicore support
cl=makeCluster(6)
registerDoParallel(cl)

# Measure run time
ptm <- proc.time()

# Load parameters for model
source("4_parameters.R")

### Scenario_1 with normal errors and weaker interviewer effects ###

# Interviewer effect
# Weaker interviewer effect with 0.25
theta10<- 0.25

# Scenario number
scenario=1

# Run simulation
source("2_iteration.R")

# Obtain best fit copula
source("3_best_fit.R")

# Generate results table
source("5_results.R")

results_1=results

# Remove results
try(rm(results))
```

```r
# Optionally save results for scenario 1
save(results_1, file="normal_1.Rdata")

### Scenario_2 with normal errors and stronger interviewer effects ###

# Interviewer effect
# Stronger interviewer effect with 0.5
theta10<- 0.5

# Scenario number
scenario=2

# Run simulation
source("2_iteration.R")

# Obtain best fit copula
source("3_best_fit.R")

# Generate results table
source("5_results.R")

results_2=results

# Remove results
try(rm(results))

# Optionally save results for scenario 2
save(results_2, file="normal_2.Rdata")

### Scenario_3 with Student-t errors and weaker interviewer effects ###

# Interviewer effect
# Stronger interviewer effect with 0.25
theta10<- 0.25

# Scenario number
scenario=3

# Run simulation
source("2_iteration.R")

# Obtain best fit copula
source("3_best_fit.R")

# Generate results table
source("5_results.R")

results_3=results

# Remove results
```

```r
try(rm(results))

# Optionally save results for scenario 3
save(results_3, file="student_1.Rdata")

### Scenario_4 with Student-t errors and stronger interviewer effects ###

# Interviewer effect
# Stronger interviewer effect with 0.5
theta10<- 0.5

# Scenario number
scenario=4

# Run simulation
source("2_iteration.R")

# Obtain best fit copula
source("3_best_fit.R")

# Generate results table
source("5_results.R")

results_4=results

# Remove results
try(rm(results))

# Optionally save results for scenario 4
save(results_4, file="student_2.Rdata")

### Scenario_5 with normal cubed errors and weaker interviewer effects ###

# Interviewer effect
# Weaker interviewer effect with 0.25
theta10<- 0.25

# Scenario number
scenario=5

# Run simulation
source("2_iteration.R")

# Obtain best fit copula
source("3_best_fit.R")

# Generate results table
source("5_results.R")

results_5=results
```

```
# Remove results
try(rm(results))

# Optionally save results for scenario 5
save(results_5, file="normal_cubed_1.Rdata")

### Scenario_6 with normal cubed errors and stronger interviewer effects ###

# Interviewer effect
# Weaker interviewer effect with 0.5
theta10<- 0.5

# Scenario number
scenario=6

# Run simulation
source("2_iteration.R")

# Obtain best fit copula
source("3_best_fit.R")

# Generate results table
source("5_results.R")

results_6=results

# Remove results
try(rm(results))

# Optionally save results for scenario 6
save(results_6, file="normal_cubed_2.Rdata")

### Scenario_7 with Clayton 270 errors and weaker interviewer effects ###

# Interviewer effect
# Weaker interviewer effect with 0.25
theta10<- 0.25

# Scenario number
scenario=7

# Run simulation
source("2_iteration.R")

# Obtain best fit copula
source("3_best_fit.R")

# Generate results table
source("5_results.R")

results_7=results
```

```r
# Remove results
try(rm(results))

# Optionally save results for scenario 7
save(results_7, file="clayton_1.Rdata")

### Scenario_8 with Clayton 270 errors and stronger interviewer effects ###

# Interviewer effect
# Stronger interviewer effect with 0.5
theta10<- 0.5

# Scenario number
scenario=8

# Run simulation
source("2_iteration.R")

# Obtain best fit copula
source("3_best_fit.R")

# Generate results table
source("5_results.R")

results_8=results

# Remove results
try(rm(results))

# Optionally save results for scenario 8
save(results_8, file="clayton_2.Rdata")

results_all=as.data.frame(rbind(results_1,results_2,results_3,results_4,results_5,
                                results_6,results_7,results_8))

proc.time() - ptm

# Summary tables
source("6_tables.R")
```

2.2 File 2: 2_iteration.R

```
# File 2/10
# Filename: 2_iteration.R
# R file for implementing the simulation iterations

# Iterate over replications
# Set seed for reproducibility with doRNG package
results<-foreach(i=1:reps, .combine='rbind', .options.RNG=123, .packages=c("copula",
                  "CDVine", "mice", "aod", "mvtnorm", "SemiParBIVProbit")) %dorng% {

  # Empty results matrix
  results0 <- rep(NA,43)
  results0[43] <-scenario

  ### 1. First generate age and urban/rural structure ###
  ########################################################

  # Generate random uniform variable to bulid age structure
  data<-data.frame(ID=1:N,u=runif(N)*100)

  # Categorical age variable matching age bins in the Zambia data
  data$agecat<-ifelse(data$u<=21.68,0, ifelse(data$u<=38.14, 2, ifelse(data$u<=53.37,
    3, ifelse(data$u<=67.78, 4, ifelse(data$u<=79.07, 5, ifelse(data$u<=86.33,6,
    ifelse(data$u<=92.44, 7, ifelse(data$u<=96.87, 8, ifelse(data$u<=100, 9)))))))))

  data$agecat<-factor(data$agecat, labels=c("15-19","20-24","25-29","30-34","35-39",
                                            "40-44","45-49","50-54","55-59"))

  # Create urban/rural variable to be 57% rural
  data$rural<-rbinom(N, 1, prob=.57)

  # Create interviewer ID with 30 interviewers to match Zambia data
  data$interviewID<-floor(runif(N, 1, numbergroups))

  # Create effectiveness for interviewers drawn from a normal with mean=0, sd=1
  effectdata<-data.frame(interviewID=1:numbergroups, effect=rnorm(numbergroups,
                                                                  mean=0, sd=1))
  data2<-merge(data, effectdata, by="interviewID")

  ### 2. Create error terms for selection and HIV equations ###
  #############################################################

  if (scenario==1) source("7_normal_errors.R")
  if (scenario==2) source("7_normal_errors.R")
  if (scenario==3) source("8_student_errors.R")
  if (scenario==4) source("8_student_errors.R")
  if (scenario==5) source("9_normal_cubed_errors.R")
  if (scenario==6) source("9_normal_cubed_errors.R")
  if (scenario==7) source("10_clayton_errors.R")
  if (scenario==8) source("10_clayton_errors.R")
```

```
  data3<-cbind(data2, errors)

### 3. Latent selection variable based on characteristics + parameter values ###
##############################################################################

# Selection equation: s^* = w + theta1*(rural) + theta2-9*age_cat +
# theta10*(interviewer effectiveness) + u_select

# Latent selection variable s_k is linear prediction
data3$s_k<- w +
  theta1*data3$rural +
  theta2*as.numeric(data3$agecat=="20-24") +
  theta3*as.numeric(data3$agecat=="25-29") +
  theta4*as.numeric(data3$agecat=="30-34") +
  theta5*as.numeric(data3$agecat=="35-39") +
  theta6*as.numeric(data3$agecat=="40-44") +
  theta7*as.numeric(data3$agecat=="45-59") +
  theta8*as.numeric(data3$agecat=="50-54") +
  theta9*as.numeric(data3$agecat=="55-59") +
  theta10*data3$effect + data3$u_select

# Probability of selection is then pnorm(s_k)
# Binary variable indicating selection s=1 if s_k>0

data3$s<-as.numeric(data3$s_k>0)

results0[1]=mean(data3$s)

# Regression of selection on covariates
reg1=lm(s~as.factor(agecat)+rural+as.factor(interviewID), data=data3)

# Wald and F tests of interviewer coefficients

vR <- rep(0,38)
chi2=wald.test(b=coef(reg1)-vR, Sigma = vcov(reg1), Terms=c(11,12,13,14,15,16,17,18,
        19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38))$result$chi2[1]

df=wald.test(b=coef(reg1)-vR, Sigma = vcov(reg1), Terms=c(11,12,13,14,15,16,17,18,
        19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38))$result$chi2[2]

f=chi2/df

results0[2]=f

### 4. Latent hiv variable based on characteristics + parameter values ###
##########################################################################

# HIV equation: hiv^* = lambda + delta1*(rural) + delta2-9*age_cat
# + delta10*(interviewer effectiveness) + u_hiv
```

```
  # Latent hiv variable hiv_k is linear prediction
  data3$hiv_k<-lambda +
    delta1*data3$rural +
    delta2*as.numeric(data3$agecat=="20-24") +
    delta3*as.numeric(data3$agecat=="25-29") +
    delta4*as.numeric(data3$agecat=="30-34") +
    delta5*as.numeric(data3$agecat=="35-39") +
    delta6*as.numeric(data3$agecat=="40-44") +
    delta7*as.numeric(data3$agecat=="45-59") +
    delta8*as.numeric(data3$agecat=="50-54") +
    delta9*as.numeric(data3$agecat=="55-59") +
    data3$u_hiv

  # Probability of HIV+ is then pnorm(hiv_k)
  # Binary variable indicating selection hiv=1 if hiv_k>0
  data3$hiv<-as.numeric(data3$hiv_k>0)

  results0[3]=mean(data3$hiv)

  ### 5. New HIV variable hiv_s reflecting selection ###
  ######################################################

  data3$hiv_s=data3$hiv
  data3$hiv_s[data3$s==0]=NA

  results0[4]=mean(data3$hiv_s, na.rm=T)

  ### 6. Imputate HIV Status Based on Age and Rural ###
  ######################################################

  data4=subset(data3, select=c(rural,agecat,hiv_s))

  mi=mice(data4, m=1)
  fit=with(mi, mean(hiv_s))

  results0[5] <- mean(sapply(1, function(x) complete(mi,x)$hiv_s))

  ### 7. Implementation of Selection Model ###
  ############################################

data3$interviewID=as.factor(data3$interviewID)

try(normal<-SemiParBIVProbit(list(s~rural+agecat+interviewID, hiv_s~rural+agecat),
                             data=data3, BivD="N", Model="BSS"))

try(frank<-SemiParBIVProbit(list(s~rural+agecat+interviewID, hiv_s~rural+agecat),
                            data=data3, BivD="F", Model="BSS"))

try(t3<-SemiParBIVProbit(list(s~rural+agecat+interviewID, hiv_s~rural+agecat),
                         data=data3, BivD="T", nu=3, Model="BSS"))
```

```
try(c90<-SemiParBIVProbit(list(s~rural+agecat+interviewID, hiv_s~rural+agecat),
                          data=data3, BivD="C90", Model="BSS"))

try(c270<-SemiParBIVProbit(list(s~rural+agecat+interviewID, hiv_s~rural+agecat),
                           data=data3, BivD="C270", Model="BSS"))

try(j90<-SemiParBIVProbit(list(s~rural+agecat+interviewID, hiv_s~rural+agecat),
                          data=data3, BivD="J90" , Model="BSS"))

try(j270<-SemiParBIVProbit(list(s~rural+agecat+interviewID, hiv_s~rural+agecat),
                           data=data3, BivD="J270", Model="BSS"))

try(g90<-SemiParBIVProbit(list(s~rural+agecat+interviewID, hiv_s~rural+agecat),
                          data=data3, BivD="G90" , Model="BSS"))

try(g270<-SemiParBIVProbit(list(s~rural+agecat+interviewID, hiv_s~rural+agecat),
                           data=data3, BivD="G270", Model="BSS"))

  ### 8. Calculate HIV Prevalence ###
  ###################################

  # Normal
  results0[6] <-try(summary(normal)$rho)

  results0[7] <-try(tableRes_0(normal)[1])
  results0[8] <-try(tableRes_0(normal)[4])
  results0[9] <-try(BIC(normal))
  results0[10] <-try(est.prev(normal)$res[2])

  # Frank
  results0[11] <-try(tableRes_0(frank)[1])
  results0[12] <-try(tableRes_0(frank)[4])
  results0[13] <-try(BIC(frank))
  results0[14] <-try(est.prev(frank)$res[2])

  # T3
  results0[15] <-try(tableRes_0(t3)[1])
  results0[16] <-try(tableRes_0(t3)[4])
  results0[17] <-try(BIC(t3))
  results0[18] <-try(est.prev(t3)$res[2])

  # c90
  results0[19] <-try(tableRes_0(c90)[1])
  results0[20] <-try(tableRes_0(c90)[4])
  results0[21] <-try(BIC(c90))
  results0[22] <-try(est.prev(c90)$res[2])

  # C270
  results0[23] <-try(tableRes_0(c270)[1])
  results0[24] <-try(tableRes_0(c270)[4])
  results0[25] <-try(BIC(c270))
```

```
  results0[26] <-try(est.prev(c270)$res[2])

  # J90
  results0[27] <-try(tableRes_0(j90)[1])
  results0[28] <-try(tableRes_0(j90)[4])
  results0[29] <-try(BIC(j90))
  results0[30] <-try(est.prev(j90)$res[2])

  # J270
  results0[31] <-try(tableRes_0(j270)[1])
  results0[32] <-try(tableRes_0(j270)[4])
  results0[33] <-try(BIC(j270))
  results0[34] <-try(est.prev(j270)$res[2])

  # G90
  results0[35] <-try(tableRes_0(g90)[1])
  results0[36] <-try(tableRes_0(g90)[4])
  results0[37] <-try(BIC(g90))
  results0[38] <-try(est.prev(g90)$res[2])

  # G270
  results0[39] <-try(tableRes_0(g270)[1])
  results0[40] <-try(tableRes_0(g270)[4])
  results0[41] <-try(BIC(g270))
  results0[42] <-try(est.prev(g270)$res[2])

  # Remove objects to prevent use after non-convergence
  rm(data, data2, data3, data4, effectdata, errors, chi2, df, f,reg1, fit, mi)

  try(rm(normal))
  try(rm(frank))
  try(rm(t3))
  try(rm(c90))
  try(rm(c270))
  try(rm(j90))
  try(rm(j270))
  try(rm(g90))
  try(rm(g270))

  results0=as.numeric(as.character(results0))

  return(results0)

}

row.names(results)=NULL

results <- data.frame(results)

names(results)=c("Consent Rate", "Interviewer F Test", "True HIV Prevalence",
  "Selected HIV Prevalence", "Imputed HIV Prevalence", "Normal RHO", "Normal Tau",
```

```
  "Normal AIC", "Normal BIC", "Normal Model HIV Prevalence", "F Tau", "F AIC",
  "F BIC", "F Model HIV Prevalence", "T Tau", "T AIC", "T BIC",
  "T Model HIV Prevalence","C90 Tau", "C90 AIC", "C90 BIC", "C90 Model HIV Prevalence",
  "C270 Tau", "C270 AIC", "C270 BIC", "C270 Model HIV Prevalence", "J90 Tau",
  "J90 AIC", "J90 BIC", "J90 Model HIV Prevalence","J270 Tau", "J270 AIC",
  "J270 BIC", "J270 Model HIV Prevalence","G90 Tau", "G90 AIC", "G90 BIC",
  "G90 Model HIV Prevalence", "G270 Tau", "G270 AIC", "G270 BIC",
  "G270 Model HIV Prevalence", "Simulation Type")
```

2.3 File 3: 3_best_fit.R

```
# File 3/10
# Filename: 3_best_fit.R
# R file for extracting the copula model with the best fit according to the AIC

# Obtain Best Fit Copula
names(results)=c("consent","int_f", "hiv", "hiv_s","hiv_imp", "rho", "tau", "aic_n",
    "bic_n", "hiv_n", "tau_f", "aic_f", "bic_f", "hiv_f","tau_t", "aic_t", "bic_t",
    "hiv_t", "tau_c90", "aic_c90", "bic_c90", "hiv_c90","tau_c270", "aic_c270",
    "bic_c270", "hiv_c270", "tau_j90", "aic_j90", "bic_j90", "hiv_j90","tau_j270",
    "aic_j270", "bic_j270", "hiv_j270", "tau_g90", "aic_g90", "bic_g90", "hiv_g90",
    "tau_g270", "aic_g270", "bic_g270", "hiv_g270", "sim_type")

results.mini<-results[,c("aic_n","aic_f","aic_t", "aic_c90", "aic_c270", "aic_j90",
                         "aic_j270", "aic_g90", "aic_g270")]

# Correct for Missing Values
results.mini[[1]][is.na(results.mini[[1]])]=999

# Best Fit is model with lowest AIC
results.mini$col.min<-apply(results.mini, 1, which.min)

results.hiv<-results[,c("hiv_n", "hiv_f", "hiv_t", "hiv_c90", "hiv_c270", "hiv_j90",
                        "hiv_j270", "hiv_g90", "hiv_g270")]

results.hiv<-cbind(results.hiv, results.mini$col.min)

names(results.hiv)[10]="min"

for(i in 1:reps) {
  results.hiv$best[i]<-results.hiv[i,results.hiv$min[i]]
}

results<-cbind(results, results.hiv$best)

results[[44]][results[[44]]==999]=NA

names(results)=c("consent","int_f", "hiv", "hiv_s","hiv_imp", "rho", "tau", "aic_n",
                 "bic_n", "hiv_n", "tau_f", "aic_f", "bic_f", "hiv_f","tau_t",
                 "aic_t", "bic_t", "hiv_t", "tau_c90", "aic_c90", "bic_c90", "hiv_c90",
                 "tau_c270", "aic_c270", "bic_c270", "hiv_c270", "tau_j90", "aic_j90",
                 "bic_j90", "hiv_j90","tau_j270", "aic_j270", "bic_j270", "hiv_j270",
                 "tau_g90", "aic_g90", "bic_g90", "hiv_g90","tau_g270", "aic_g270",
                 "bic_g270", "hiv_g270", "sim_type", "hiv_best")
```

2.4 File 4: 4_parameters.R

```
# File 4/10
# Filename: 4_parameters.R
# R file for loading simulation parameter values

# Number of replications
reps=1000

# Set up dataset N
N<-6500

# Create interviewer ID with 30 interviewers
numbergroups<-30

# Regression parameters were obtained from a bivariate probit model for
# Zambia DHS 2007 (Men)

# Parameters for Consent Equation

# Constant
w<-.604

# Rural effect
theta1<- .123

# Parameters for age category effects
theta2<- -.025
theta3<- -.023
theta4<-  .031
theta5<-  .101
theta6<-  .148
theta7<-  .047
theta8<-  .035
theta9<-  .154

# Set Parameters for HIV Equation

# Constant
lambda<- -1.156

# Rural Effect
delta1<- -.396

# Parameters for age category effects
delta2<- .229
delta3<- .703
delta4<-1.036
delta5<-1.147
delta6<-1.203
delta7<-1.063
```

```r
delta8<- .834
delta9<- .661

### Summary Table Function
tableRes_0 <- function(object1){

  summ <- summary(object1)

  round(cbind(summ$KeT,summ$CIkt[1],summ$CIkt[2],
              AIC(object1) ),2)

}
```

2.5 File 5: 5_results.R

```
# File 5/10
# Filename: 5_results.R
# R file for extracting results

# Calcualte mean proportional error and root mean square error

# Error
results$hiv_n1_e=(results$hiv_n-results$hiv)
results$hiv_imp1_e=(results$hiv_imp-results$hiv)
results$hiv_best1_e=(results$hiv_best-results$hiv)
results$hiv_c2701_e=(results$hiv_c270-results$hiv)
results$hiv_t1_e=(results$hiv_t-results$hiv)

# Proportional Error
results$hiv_n1=(results$hiv_n-results$hiv)/results$hiv
results$hiv_imp1=(results$hiv_imp-results$hiv)/results$hiv
results$hiv_best1=(results$hiv_best-results$hiv)/results$hiv
results$hiv_c2701=(results$hiv_c270-results$hiv)/results$hiv
results$hiv_t1=(results$hiv_t-results$hiv)/results$hiv

# RMSE
# Root Mean Square Error
results$hiv_n1_rmse=sqrt(mean(results$hiv_n1_e^2, na.rm=T))
results$hiv_imp1_rmse=sqrt(mean(results$hiv_imp1_e^2, na.rm=T))
results$hiv_best1_rmse=sqrt(mean(results$hiv_best1_e^2, na.rm=T))
results$hiv_c2701_rmse=sqrt(mean(results$hiv_c2701_e^2, na.rm=T))
results$hiv_t1_rmse=sqrt(mean(results$hiv_t1_e^2, na.rm=T))
```

2.6 File 6: 6_tables.R

```
# File 6/10
# Filename: 6_tables.R
# R file for preparing tables


# Results tables

# New variable for copula model
# Proportional Error
results_all$hiv_copula1=NA

results_all$hiv_copula1[results_all$sim_type %in% 1]=
  results_all$hiv_best1[results_all$sim_type %in% 1]

results_all$hiv_copula1[results_all$sim_type %in% 2]=
  results_all$hiv_best1[results_all$sim_type %in% 2]

results_all$hiv_copula1[results_all$sim_type %in% 3]=
  results_all$hiv_t1[results_all$sim_type %in% 3]

results_all$hiv_copula1[results_all$sim_type %in% 4]=
  results_all$hiv_t1[results_all$sim_type %in% 4]

results_all$hiv_copula1[results_all$sim_type %in% 5]=
  results_all$hiv_best1[results_all$sim_type %in% 5]

results_all$hiv_copula1[results_all$sim_type %in% 6]=
  results_all$hiv_best1[results_all$sim_type %in% 6]

results_all$hiv_copula1[results_all$sim_type %in% 7]=
  results_all$hiv_c2701[results_all$sim_type %in% 7]

results_all$hiv_copula1[results_all$sim_type %in% 8]=
  results_all$hiv_c2701[results_all$sim_type %in% 8]

# Proportional Error
results_all$hiv_copula_rmse=NA

results_all$hiv_copula_rmse[results_all$sim_type %in% 1]=
  results_all$hiv_best1_rmse[results_all$sim_type %in% 1]

results_all$hiv_copula_rmse[results_all$sim_type %in% 2]=
  results_all$hiv_best1_rmse[results_all$sim_type %in% 2]

results_all$hiv_copula_rmse[results_all$sim_type %in% 3]=
  results_all$hiv_t1_rmse[results_all$sim_type %in% 3]

results_all$hiv_copula_rmse[results_all$sim_type %in% 4]=
  results_all$hiv_t1_rmse[results_all$sim_type %in% 4]
```

```
results_all$hiv_copula_rmse[results_all$sim_type %in% 5]=
  results_all$hiv_best1_rmse[results_all$sim_type %in% 5]

results_all$hiv_copula_rmse[results_all$sim_type %in% 6]=
  results_all$hiv_best1_rmse[results_all$sim_type %in% 6]

results_all$hiv_copula_rmse[results_all$sim_type %in% 7]=
  results_all$hiv_c2701_rmse[results_all$sim_type %in% 7]

results_all$hiv_copula_rmse[results_all$sim_type %in% 8]=
  results_all$hiv_c2701_rmse[results_all$sim_type %in% 8]

# Table 1 Summary Statistics

table1=cbind(aggregate(results_all$hiv,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE),
             aggregate(results_all$hiv_s,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE),
             aggregate(results_all$consent,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE),
             aggregate(results_all$int_f,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE))

table1=round(table1[c(2,4,6,8)],3)

names(table1)=c("True HIV Prevalence (%)", "Observed HIV Prevalence (%)",
                "Consent (%)", "Interviewer F")

rownames(table1)=c("Normal: Weaker Interviewer Effects",
                   "Normal: Stronger Interviewer Effects",
"Student-t: Weaker Interviewer Effects", "Student-t: Stronger Interviewer Effects",
"Normal Cubed: Weaker Interviewer Effects",
"Normal Cubed: Stronger Interviewer Effects", "Clayton-270: Weaker Interviewer Effects",
"Clayton-270: Stronger Interviewer Effects")

# Table 2 Mean Proportional Error

table2=cbind(aggregate(results_all$hiv_imp1,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE),
             aggregate(results_all$hiv_n1,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE),
             aggregate(results_all$hiv_copula1,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE))

table2=round(table2[c(2,4,6)],3)

names(table2)=c("Imputation Model", "Standard (Gaussian) Selection Model",
                "Copula Selection Model")

rownames(table2)=c("Normal: Weaker Interviewer Effects",
```

```
                    "Normal: Stronger Interviewer Effects",
                    "Student-t: Weaker Interviewer Effects",
                    "Student-t: Stronger Interviewer Effects",
                    "Normal Cubed: Weaker Interviewer Effects",
                    "Normal Cubed: Stronger Interviewer Effects",
                    "Clayton-270: Weaker Interviewer Effects",
                    "Clayton-270: Stronger Interviewer Effects")

# Table 3 Root Mean Square Error

table3=cbind(aggregate(results_all$hiv_imp1_rmse,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE),
             aggregate(results_all$hiv_n1_rmse,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE),
             aggregate(results_all$hiv_copula_rmse,by=list(results_all$sim_type),
                       FUN=mean, na.rm=TRUE))

table3=round(table3[c(2,4,6)],3)

names(table3)=c("Imputation Model", "Standard (Gaussian) Selection Model",
                "Copula Selection Model")

rownames(table3)=c("Normal: Weaker Interviewer Effects",
                    "Normal: Stronger Interviewer Effects",
                    "Student-t: Weaker Interviewer Effects",
                    "Student-t: Stronger Interviewer Effects",
                    "Normal Cubed: Weaker Interviewer Effects",
                    "Normal Cubed: Stronger Interviewer Effects",
                    "Clayton-270: Weaker Interviewer Effects",
                    "Clayton-270: Stronger Interviewer Effects")
```

2.7 File 7: 7_normal_errors.R

```
# File 7/10
# Filename: 7_normal_cubed_errors.R
# R file for drawing bivariate normal cubed errors


# Matrix for Error Terms
rho=-0.75
mu1=c(0,0)
Sigma<-matrix(data=c(1, rho, rho, 1), nrow=2, ncol=2)

# Load parameters for model
source("4_parameters.R")

### Normal Errors

errors<-as.data.frame(mvrnorm(N, mu=mu1, Sigma))
names(errors)<-c("u_select", "u_hiv")
```

2.8 File 8: 8_student_errors.R

```
# File 8/10
# Filename: 8_student_errors.R
# R file for drawing errors from Student-t


# Load parameters for model
source("4_parameters.R")

### Student-t Errors

clay.cop.6 = tCopula(0.75, dim=2, df=4)
tau(clay.cop.6)
errors=as.data.frame(rCopula(N, clay.cop.6))

names(errors)<-c("u_select", "u_hiv")

errors$u_select=errors$u_select-mean(errors$u_select)
errors$u_select=errors$u_select/sd(errors$u_select)

errors$u_hiv=errors$u_hiv-mean(errors$u_hiv)
errors$u_hiv=errors$u_hiv/sd(errors$u_hiv)

errors$u_hiv=-errors$u_hiv
```

2.9 File 9: 9_normal_cubed.R

```
# File 9/10
# Filename: 9_normal_cubed.R
# R file for drawing errors from normal cubed

# Matrix for Error Terms
rho=-0.75
mu1=c(0,0)
Sigma<-matrix(data=c(1, rho, rho, 1), nrow=2, ncol=2)

# Load parameters for model
source("4_parameters.R")

### Normal Cubed Errors

errors<-as.data.frame(mvrnorm(N, mu=mu1, Sigma))
names(errors)<-c("u_select", "u_hiv")

# Replace Errors with Errors Squared
errors$u_select=(errors$u_select^3)

errors$u_hiv=(errors$u_hiv^3)
```

2.10 File 10: 10_clayton_errors.R

```
# File 10/10
# Filename: 10_clayton_errors.R
# R file for drawing errors from Clayton 270

# Load parameters for model
source("4_parameters.R")

# Clayton Errors
clay.cop.6 = archmCopula(family="clayton", dim=2, param=4)
tau(clay.cop.6)
errors=as.data.frame(rCopula(N, clay.cop.6))

names(errors)<-c("u_select", "u_hiv")

errors$u_select=errors$u_select-mean(errors$u_select)
errors$u_select=errors$u_select/sd(errors$u_select)

errors$u_hiv=errors$u_hiv-mean(errors$u_hiv)
errors$u_hiv=errors$u_hiv/sd(errors$u_hiv)

errors$u_select=-errors$u_select
```

3 Code for Figure 2 (Drawing from Copulae)

```
# Copula plots for Figure 2 in the main text
library(VineCopula)

mar=c(3,3,3,7)
par(mfrow = c(2, 2),mar=c(3.1,4.1,3.1,5.1),cex.main = 1.7, cex.lab = 1, cex.axis = 1)

nf <- layout(matrix(c(1,1,2,2,3,3,4,4,5,5,6,6,0,7,7,0), 4, 4, byrow=TRUE),
             respect=FALSE)

dat = BiCopSim(N = 1000, family = 5, par =BiCopTau2Par(family = 5,
                                                       tau = -0.5))
BiCopMetaContour(u1 = dat[, 1], u2 = dat[, 2], bw = 1, size = 100, levels = c(0.01,
   0.05, 0.1, 0.15, 0.2), family = 5, par = BiCopTau2Par(family = 5,

                 main = "Frank")

dat = BiCopSim(N = 1000, family = 2, par =BiCopTau2Par(family = 1,
                                                       tau = -0.5), par2=3)
BiCopMetaContour(u1 = dat[, 1], u2 = dat[, 2], bw = 1, size = 100, levels = c(0.01,
 0.05, 0.1, 0.15, 0.2), family = 2, par = BiCopTau2Par(family = 1,
                            tau = -0.5), par2=3,margins="norm", main = "Student-t")

dat = BiCopSim(N = 1000, family = 23, par = BiCopTau2Par(family = 23,
                                                         tau = -0.5))
BiCopMetaContour(u1 = dat[, 1], u2 = dat[, 2], bw = 1, size = 100, levels = c(0.01,
  0.05, 0.1, 0.15, 0.2), family = 23, par = BiCopTau2Par(family = 23,
 tau = -0.5), margins="norm", main = "Clayton 90 degrees")

dat = BiCopSim(N = 1000, family = 33, par =BiCopTau2Par(family = 33,
                                                       tau = -0.5))
BiCopMetaContour(u1 = dat[, 1], u2 = dat[, 2], bw = 1, size = 100, levels = c(0.01,
 0.05, 0.1, 0.15, 0.2), family = 33, par = BiCopTau2Par(family = 33,
 tau = -0.5), margins="norm",
                 main = "Clayton 270 degrees")

dat = BiCopSim(N = 1000, family = 24, par =BiCopTau2Par(family = 24,
                                                       tau = -0.5))
BiCopMetaContour(u1 = dat[, 1], u2 = dat[, 2], bw = 1, size = 100, levels = c(0.01,
   0.05, 0.1, 0.15, 0.2), family = 24, par = BiCopTau2Par(family = 24,
tau = -0.5), margins="norm",
                 main = "Gumbel 90 degrees")

dat = BiCopSim(N = 1000, family = 34, par =BiCopTau2Par(family = 34,
                                                       tau = -0.5))
BiCopMetaContour(u1 = dat[, 1], u2 = dat[, 2], bw = 1, size = 100, levels = c(0.01,
  0.05, 0.1, 0.15, 0.2), family = 34, par = BiCopTau2Par(family = 34,
tau = -0.5), margins="norm",
                 main = "Gumbel 270 degrees")
```

```
dat = BiCopSim(N = 1000, family = 1, par =BiCopTau2Par(family = 1, tau = -0.5))
BiCopMetaContour(u1 = dat[, 1], u2 = dat[, 2], bw = 1, size = 100, levels = c(0.01,
 0.05, 0.1, 0.15, 0.2), family = 1, par = BiCopTau2Par(family = 1, tau = -0.5),
 margins="norm", main = "Gaussian")
```